special-topic
bulletin from

DEPARTMENT OF
EDUCATIONAL AND
PSYCHOLOGICAL RESEARCH

MALMÖ SCHOOL OF EDUCATION
LUND UNIVERSITY

didakometry

Bierschenk, I.:

AN INFORMATION PROCESSING EXPERIMENT

No. 62 May 1981

AN INFORMATION PROCESSING EXPERIMENT

A Method for the Generation of an Intermediate Language for Representing Scientific Information

Inger Bierschenk

Bierschenk, I. An information processing experiment. A method for the generation of an intermediate language for representing scientific information. Didakometry (Malmö, Sweden: Department of Educational & Psychological Research), No. 62, 1981.

This experiment starts with the assumption that the structure and representation of scientific information should correspond to the cognitive structure assumed to exist in both user and producer of information. The model of investigation of cognitive representation is based on overt manifestations of concepts and conceptual relations as they emerge in the abstract language of titles of scientific documents. On the basis of this language structure an algorithm has been developed and tested using the cue function of prepositions for automatic conceptual coding. The relevance of the concepts is judged with respect to a schema model containing some basic components of research itself. The algorithm generates the assumed scientific concepts and assigns them to different data registers.

Keywords: Cognitive science, computational linguistics, information processing, information science, linguistics, representation language, text processing, thesaurus.

To Bernhard

CONTENTS

Preface

This monograph presents a model, interdisciplinary conceived and linguistically oriented, which can be used for the study of information processing. The notion "information" is central for studies within, for example, information science, artificial intelligence, computational linguistics and the emerging field of cognitive science. It is used in the study of mechanisms that make possible the representation of information derived from symbols. The computation of language utilizes the computer's capacity of being a symbol handling machine. Therefore, I believe that a linguistically oriented model has to be incorporated in the information processing of natural language. Otherwise these attempts will not easily come up with meaningful results.

The strategy adapted here is based on a few, but quite general, principles. I believe that these principles can be used in the procedure of mediating information between man and the machine. If the interpretation of language data is based on cognitive functions such as they are studied by cognition oriented scientists, much of the confusion between the outcome based on formal logic and the logic of conception may disappear. After all, theories developed by computer scientists and linguists oriented towards pure logical approaches have not been very successful in explaining information processing based on human language.

What I hope to show is that a certain type of language within a particular context can be studied computationally and provide meaningful results when criteria for the interpretation of information are set up in advance. The experiment in information processing carried out and presented in this work show that much of the complexity (ambiguity) involved in natural language processing diminishes.

An approach of this kind uses theoretical strands from several scientific disciplines. I think that the strategy taken will open up new perspectives.

I have learnt a great deal about the processing of written text as opposed to scattered sentences cited by generations of philosophers, linguists and psychologists through many years of theoretical and empirical research at the Department of Educational and Psychological Research, University of Lund-Malmö, Sweden. Many of the key ideas have been developed within several research programmes that have had their financial support from the National Board of Education. Although the main phases of this research have been performed in Malmö, I have received substantial assistance and support from researchers elsewhere. I would like to take this opportunity to express my indebtedness to the following persons.

It is a pleasure for me to convey my deep appreciation to Professor Sture Allén at the Department of Computational Linguistics, University of Göteborg. I wish to thank him for his constructive criticism and his valuable suggestions for improvements based on several drafts. I am convinced that without his interest in my work, my plans and the completion of my study could have materialized only with considerable difficulty.

I also would like to express my gratitude to fellow researchers in Sweden with whom I have had the possibility to discuss my work in seminars on linguistics, computational linguistics and information science. Many interesting points of discussion and concrete suggestions have been of valuable help. I especially wish to thank Dr. Benny Brodda at the Department of Linguistics, University of Stockholm. His reading and commenting on an earlier version, presented as doctoral dissertation in 1980, has been very helpful for the further elaboration of the theoretical discussion. In this connection I also want to thank Mats Eeg-Olofsson, Fil. Kand., and Dr. Bo Ralph at the Department of Computational Linguistics, University of Göteborg for important comments.

Several arguments presented here have been developed as a response to discussions held with scientists whom I had the pleasure to meet at a Nobel Symposium on Text Processing in Stockholm 1980. For their kind listening and their informative questions I especially wish to thank Professors Walter Kintsch and Erik Sandewall.

An extensive part of the manuscript was developed during a visit to Canada during the academic year 1979-80. I would like to take this opportunity to thank a number of colleagues there who have stimulated me in several ways. Dr. Leo Mos at the Center for Advanced Study in Theoretical Psychology made it possible for me to listen and learn, and also to discuss research problems in formal seminars. Dr. Michael Katzko at the Center has carefully reviewed the manuscript from the point of view of cognitive psychology. His valuable advice concerning some theoretical aspects has been of great importance to me.

My thanks also concern Canadian researchers in computer science. I am very grateful to Professor Lenhart Schubert, who let me follow his interesting course in artificial intelligence during the fall semester 1979. I have also acquired valuable insights into the wide field of data management through discussions with Professor Nick Cercone and Randy Goebel, M.Sc., in Vancouver.

The experiments I present here would not have been successful without the kind assistance of several persons at several stages of development. The translation of the proposed algorithm into a workable computer program was carried out by Agneta Sternerup-Hansson, Fil. Kand., at the Computing Centre of the University of Lund-Malmö. I am grateful for the interest she has shown and for the many valuable suggestions she has made during the development of the analysis.

In computational projects, success is greatly dependent on the proper transfer of data. I wish to thank Mrs. Ingegerd Johansson, Malmö, for her extensive work transferring the material onto magnetic tape.

I am also aware of the great importance played by the language practice through which research results are conveyed to the reader. Dr. Sölve Ohlander at the English

Department, University of Göteborg, has kindly revised the English manuscript with respect to proper wording and stylistic aspects. I wish to thank him for his discerning reading of my English. I also thank Richard Fisher, M.A., at the English Department, University of Lund-Malmö for his reading of some parts of this version.

I would also like to thank Mrs. Karin Dahlberg, Malmö, for her care and thoroughness in typing the manuscript.

Last but not least I would like to express my deep gratitude to my husband and scientific inspirer, Professor Bernhard Bierschenk, Department of Educational and Psychological Research, University of Lund-Malmö. He introduced me to the problem field with which this study is concerned. The ample opportunities I have had, over the years, of discussing various research problems with him, both on formal and informal occasions, have been an invaluable asset in my work. In the development and the carrying out of this experiment he always gave me his strong support. With these experiences alive in my memory and a hope for continuing fruitful collaboration, I dedicate this study to him.

Ljunghusen, Spring 1981
Inger Bierschenk

1. INTRODUCTION

From the point of view of information technology, society today is characterized by a high level of development. New information and documentation systems (I&D systems) have been, and are being, devised in many places around the world. But the effects of this development vary considerably. A case in point is the phenomenon which may be called "information frustration". *Information,* the first part of this compound, refers to the meaning an individual abstracts from data. *Data* are characterized by physical existence in the sense that they can be counted, measured, and classified. Thus data must be organized in such a way that information may be inferred or interpreted. The second part of the compound refers to a psychological phenomenon connected with the barriers that have emerged between the information searcher and the goal of the search, namely to get access to the information contained in the documents.

One of the more serious barriers in this connection is the multiplicity of document descriptions, which are disseminated without the individual's having access to the documents themselves. This circumstance creates an often well-grounded uncertainty about the real existence of a document. Another barrier is that advanced technical systems have been built up as a connecting link between the information searcher and the information potentially available. The user is often afraid of the machines, and he is not always aware of the logic on which the programs are based. Another complicating factor is that the "troublesome" routines are not free of charge; further, the systems usually do not provide user-oriented tutoring functions. An example of a desirable guiding program is "Interactive consulting via natural language", developed by Shapiro & Kwasny (1975). Large groups of potential users are in effect prevented access to such information as is absolutely necessary for the development and management of information. A third barrier is that as yet no computer programs have been developed which can handle document descriptions presented in several different languages. Such programs would be of great help in international cooperation concerning matters of I&D for different subject areas.

For some time it has been claimed that modern society is characterized by "information overflow" and "information explosion" (Price, 1963; Anderla, 1973). Such claims, however, cannot be fully accepted, since objections may be raised against Price's global methods and overgeneralizations (see Gilbert, 1978; B. Bierschenk, 1979). Moreover, several studies have shown that many professional categories requiring information, such as researchers and teachers (B. Bierschenk, 1974; Jernryd, 1976; 1978), seem to be suffering from a considerable lack of information. The claims obviously refer to a flow of data or documents (see below) with which the individual tends to feel increasingly unfamiliar, resulting in an inability to abstract a meaning (information) out of the flow.

To set things right, increased research on the basic mechanisms at work in information selection, storage and retrieval is required. Increased communication is no solution to the problem. There is a need for greater insights into and a better understanding of how information can be made available, enabling the individual to adapt it to the cognitive structures characterizing the models which govern his search for information. In the present work the available information consists of so-called "non fugitive information", i.e. the kind of information that has been documented. A *document* may then be defined as

> a written or printed record, being the definite proof that information exists.

Thus *documentation* means, among other things, the supplying of documents. Libraries traditionally supply documents. But when it comes to the computer-based I&D systems, the situation is different. They traditionally supply only document descriptions or references to documents. Thus documentation has also come to imply the organization and representation of document descriptions. Against this background the concept of *data base*, in connection with information and documentation, may be defined as

> a collection of data which is part of another collection of data (documents) and which consists of at least one register, which is organized in such a way that its structure is suited for a precise description of the documents of which the first collection of data is a part.

(This definition is a close adaptation from Wersig & Neveling, 1976, where 'collection' refers to an "organized body of stored items".) On the basis of this definition, *register* refers to a list of specified data, whose purpose is to describe a document.

The possibility of retrieving meaningful information from a data base is determined by the way in which the documents are described and how these descriptions are structured in the registers. The descriptions constitute representations of physical documents. The basic problem for every modern I&D system is to find the formats of representation most adequate to its different goals, and to develop inference mechanisms in order to make for comprehension of stored information.

One format of representation may be a purely bibliographic description. Examples of bibliographic elements are "name of author", "co-author", "title", or "name of journal". Another kind of representation may be "outward" characteristics, such as "colour of cover", "layout", "binding", or "general condition" of a document. A system may also need to discriminate between types of documents, i.e. books as distinct from non-book material, a way of characterization which has caused trouble all over the world, partly because the borderlines between what should and what should not be regarded as books have become increasingly blurred due to new composing and copying techniques.

All these ways of characterizing a document may be said to be descriptive. Many information systems are based on an organization of such descriptive data about documents, and the retrieval from a search in such data bases should, therefore, be called document retrieval.

The use of the term "information retrieval" presupposes that the representation is a result of some form of *analysis* of what a document is intended to communicate. This process is referred to by many terms, some of which are "content analysis", "content detection" and "making judgement of aboutness" (see e.g. Fairthorne's discussion in the Annual Review of Information Science and Technology, 1969).

A document description based on content analysis adds a cognitive dimension to the I&D system which, in turn, may have different representational levels. The analysis may be represented by means of some keywords (or descriptors) being added to the document description, relating the document to a certain conceptual structure. *Keyword* is a word or a term which is assigned to the document from the document itself for indexing purposes. A *descriptor* is a main term or phrase which, for the same purposes, is drawn from a thesaurus (see below). In information and documentation *word* is defined as being a string of characters, while *term* refers to a word or phrase designating a *concept*. (The definitions in connection with this subject field are taken from Wersig & Neveling, 1976.)

Both keywords and descriptors may be assigned to documents on the basis of title, list of contents or search in the text. Examples of higher levels are *abstracts*, which are a kind of compressed informative or indicative summaries, and *extracts*, which here are to be regarded primarily as selections of representative paragraphs. These document descriptions are usually provided not by the author himself, but by someone else.

In devising bibliographic descriptions it is convenient to follow some international standard, e.g. the American

Psychological Association (APA). In general, the design of a bibliographic data base does not cause any difficulties when the level of structuring is low. Structuring based on content in documents, however, has raised many problems, since it is based on interpretation. The structuring principles within library science, for example, rely on philosophical thinking, as manifested in a classification system. Information is then determined in such a way that the description of a document is adapted to the structure that characterizes the classification systems. Examples of such systems are the SAB (Sveriges Allmänna Biblioteksförening/The General Library Society of Sweden) and the UDC (Universal Decimal Classification). However, in connection with computer-based I&D systems means of assistance have been developed in the form of the kind of structured dictionaries that are usually called thesauri. A *thesaurus* is characterized by an organized display of the relations which hold between terms and descriptors and which define these. Therefore, the thesaurus is used as an encoding device in the information search process: its structure determines the success of the information retrieval. No matter what organization form the designer of the system chooses, he has to see to it that information is made explicit. Thus the central problem of an information system is to find ways of representing information in such a way that the structure in which the author communicates it corresponds to the structure in which the reader perceives it or wishes it to be. Such a way of reasoning makes it hardly possible to try to solve the so-called information problem without focusing on its cognitive aspects.

This general introduction to the field of information and documentation, terminology and computer-based systems and activities connected with it, contains the ideas which have governed the present attempt to tackle the information problem. The aims of this study are summarized below.

The main goal of this research is, in short, to construct a thesaurus in which the structure shall represent the conception of a specific branch of science.

Psychological Association (APA). In general, the design of a bibliographic data base does not cause any difficulties when the level of structuring is low. Structuring based on content in documents, however, has raised many problems, since it is based on interpretation. The structuring principles within library science, for example, rely on philosophical thinking, as manifested in a classification system. Information is then determined in such a way that the description of a document is adapted to the structure that characterizes the classification systems. Examples of such systems are the SAB (Sveriges Allmänna Biblioteksförening/The General Library Society of Sweden) and the UDC (Universal Decimal Classification). However, in connection with computer-based I&D systems means of assistance have been developed in the form of the kind of structured dictionaries that are usually called thesauri. A *thesaurus* is characterized by an organized display of the relations which hold between terms and descriptors and which define these. Therefore, the thesaurus is used as an encoding device in the information search process: its structure determines the success of the information retrieval. No matter what organization form the designer of the system chooses, he has to see to it that information is made explicit. Thus the central problem of an information system is to find ways of representing information in such a way that the structure in which the author communicates it corresponds to the structure in which the reader perceives it or wishes it to be. Such a way of reasoning makes it hardly possible to try to solve the so-called information problem without focusing on its cognitive aspects.

This general introduction to the field of information and documentation, terminology and computer-based systems and activities connected with it, contains the ideas which have governed the present attempt to tackle the information problem. The aims of this study are summarized below.

The main goal of this research is, in short, to construct a thesaurus in which the structure shall represent the conception of a specific branch of science.

For this purpose an experimental data base containing document descriptions from a random sample of researchers in Sweden is used. The thesaurus is planned to display conceptual relations as they have emerged in Swedish educational research from 1937 up to now.

It is assumed that knowledge and techniques from several sciences are necessary. These include linguistics, especially computational linguistics, library science, information science, artificial intelligence and cognitive and educational psychology.

This study presents a model and a method that, in contrast to ordinary approaches in this field of application, is built on intermediary functions of language, thus hypothesizing certain language structures to coincide with cognitive ones. Such structures shall constitute the thesaurus entries and be generated from titles of the scientific documents and no other kind of text. The thesaurus is intended to build on functions instead of classification.

In order to reach the main goal, some intermediate steps have to be taken. Those <u>subgoals</u> are listed below and constitute representations of the research process itself together with additional information to the reader.

1. To present some basic principles governing the systematization of information.
   These principles are outlined in Chapters 2 and 3.
2. To present a general model for re-cognition based on such components in titles as are derived from research on cognition, and to develop an algorithm capable of coding titles in accordance with the model.
3. To analyse and demonstrate the linguistic representation format of the model, and to show the extent to which regularities in language can be used in the communication of scientific information within a particular field of application.
   This theoretical discussion is presented in Chapter 4.
4. To map structural relations in the linguistic representation format and to describe those structures quantitatively, and also to analyse the relationship between structures and types of scientific documents.
   The results of the structural analysis are given in Chapter 5.

5. To analyse and describe the components' conceptual foundation in the particular data base, and to indicate intermediate language functions.

6. To demonstrate the function of the registers in the data base and the relevance of their entries as a basis for a functionally oriented thesaurus.
   The presentation of these results is to be found in Chapter 6.
   In the last Chapter some concluding remarks are made.

2. SOME STRUCTURING PRINCIPLES OF RELEVANCE TO INFORMATION SCIENCE

Communication processes are made possible by means of systems that are open with respect to information input. An information system designed to handle documents or document descriptions should have as its primary goal to provide an overview, as comprehensive as possible, of incoming information. Its fundamental purpose, therefore, should be to create order. A universal conception of the creation of order has guided library systems in the past and still does. With the computerization of library science the possibility arose to automatically sort and organize bibliographic indexes and catalogues (e.g. author indexes). Such so-called non-intellectual routines could be taken over by the machines. Difficulties arose, however, when it came to intellectual routines, such as indexing, i.e. the process that includes analysis and classification of documents. The need for automatic generation of subject indexes brought to the fore other than bibliographic problems. These circumstances, together with the highly increased output of scientific documents of different kinds, led to the establishment of a new discipline, namely information science, which, besides library science, involves systems theory, automatic text processing, linguistics, and computer science. A fairly good picture of the content of and development within this field is given in the Annual Review of Information Science and Technology, whose first issue appeared in 1966.

The process of selection, storage and retrieval of documents and document descriptions at the bibliographic level is no longer a technical problem. This level of representation will therefore not be further considered. Instead, the intellectual routines that are related to analysis and description of document content for information retrieval will be focused on.

There are authors who, from the standpoint of natural science, argue that progress is made through detection of facts or through new ideas and events (e.g. Price, 1963). From the point of view of social science however, it may be claimed that progress is to a greater extent made through new principles of organization, new theories, new relationships ("repacking of older information" according to Anderla, 1973, p. 120). It would hardly be realistic to believe that an I&D system would be able to survey the total amount of information, especially since information is constantly changing. The adaptive properties of an I&D system, therefore, are reflected in its capacity to structure information in a flexible way. In the following section a short presentation of some principles for structuring of information will be given.

2.1 Hierarchies

For the purpose of collecting "all" documented information in libraries, the UDC (Universal Decimal Classification) was developed for the organization of book stocks on shelves. The starting point was Dewey's Decimal Classification (DC). The decimals are retained within UDC, and the literature is organized in ten main categories, designated by the numbers 0-9. The UDC differs from the DC by using more than one decimal.

The task of classification systems is to define the relationships between the single elements. The hierarchical structure that characterizes the UDC is similar to a tree structure with strict super- and subordination. *Hierarchy* is defined as a strict organizational system. By means of the ten main classes related subjects are grouped together, even though a strict ranking order in the subclasses is maintained. Consequently, restructuring is only possible through addition and expansion along the outer edges of the tree. On the other hand, the UDC system may more easily than other library systems be used for retrieval purposes (see Mølgaard-Hansen, 1968). In order to illustrate the hierarchy as a structuring principle, Box 1 shows the structure of General Linguistics as presented in the Swedish version of the UDC system (UDK, 1961).

General Linguistics is classified directly under main class 4, i.e. "Linguistics, Philology". Class 41 indicates that the subject field is rather old. It may be compared with, e.g., Psychology, which is classified in main class 1 under "Philosophy", although as number 159.9 in the tree. This means that it is a subject field that has been added to the original structure at a later stage. Psychology, however, contains six times as many branches as Linguistics.

Box 1. Example of hierarchy: General Linguistics as structured by the Swedish version of the UDC.

41	General Linguistics
411	Orthographic rules. Correct spelling. Orthography
.4	Orthographic reform
412	Word classes
413	Lexicology. Dictionaries
.1	Words according to meaning. Semasiology
.11	Place names
.13	Names of persons
.14	Homonyms and synonyms
.163	Foreign words. Loanwords
.164	Professional terms. Technical terms
.2	Dictionaries classified according to different aspects
	⋮
414	Phonetics. Phonology
415	Grammar
.4	Etymology. Semantics. Semiotics
.5	Morphology. Accidence
.6	Syntax
416	Metrics. Prosody
417	Reference sciences. Hermeneutics. Exegesis. Textual criticism
418	Original sources of Linguistics

The assignment of a document's place in a classification system depends on the preciseness of the indexing, i.e. the assignment of keywords or descriptors to the document. A way to avoid subjectivity in making decisions is to use a controlled terminology. Such a terminology for document description has been developed in computer-based I&D systems for the purpose of representing concepts and conceptual relations. Regardless of what principles may govern the structuring of a subject field, the structure employed is represented in a thesaurus. The thesaurus is an aid in both indexing and information search, which implies that its structure effects the precision of the information retrieval from the respective field of knowledge. The best-known thesaurus in the field of education is that developed by ERIC (Educational Resources Information Center). The Thesaurus of ERIC Descriptors (1975) is in principle hierarchically structured. Three types of function terms are used for structuring the vocabulary. These types are, basically, USE ("see or use") and UF ("used for"); BT ("broader term") and NT ("narrower term"); RT ("related term"). The terms in the first group have a controlling function. By USE is indicated which term is the more correct one (used by professionals), whereas UF indicates which term has been used earlier in designating the same particular field. Thus the UF function admits retrospection, i.e., contact with the historical development is maintained.

The descriptors are structured by means of the hierarchical relations BT and NT. The importance of these links is that the user can let the built-in hierarchies guide his search. The possibility of relating different hierarchies is provided by the RT function. There is also a possibility of updating, i.e., additional descriptors may be assigned to the system, depicting the progress of the subject field.

2.2 Facets

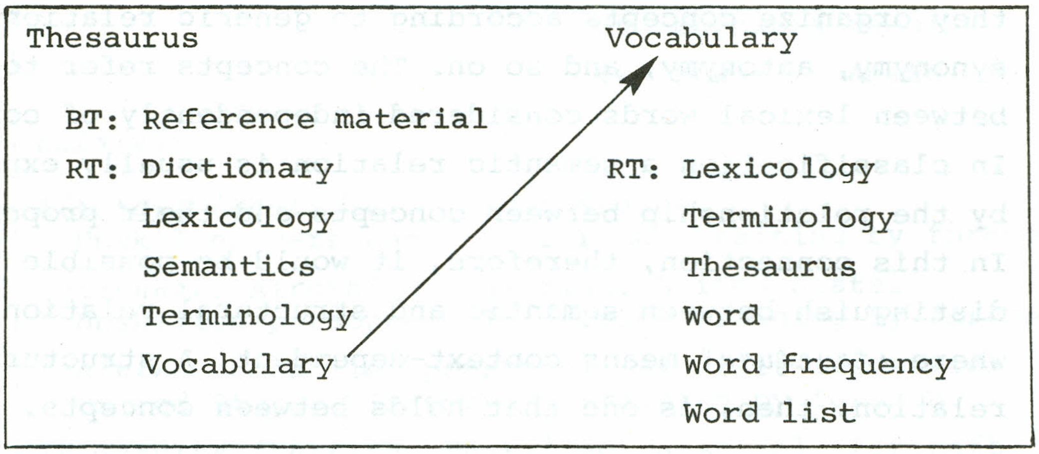
The concept of facet refers to an aspect of a document, a subject, and so on. In an analysis of facets a complex subject field is decomposed into as many aspects as possible. One of the oldest philosophical classification systems is Ranganathan's "tree of knowledge", in which the world is described by means of the five facets "Space", "Matter", "Economy", "Time", and "Personality". Ranganathan's (1964) Colon Classification has given rise to several thesauri, one of the better-known of which is the Thesaurofacet, which classifies electrotechnical engineering and related fields. This thesaurus is structured in "fundamental facets", "sub facets" and "hierarchies" (see Aitchison, 1970).

One type of facet classification within the field of information and documentation is to be found in Terminology of Documentation (Wersig & Neveling, 1976), published by UNESCO. In comparison with the ERIC Thesaurus, it may be noted that the Terminology of Documentation uses the same function terms, i.e. BT, NT and RT, according to the same principles as in ERIC. There is also a function named OT ("opposite term"). Since this thesaurus is not linked to any retrieval system, it has no reference terms (USE, UF). The Terminology of Documentation is an attempt to standardize terminology within the I&D field in the English, German, French, Spanish and Russian languages. For the classification three aims have been deemed important (p. 12), namely (1) to connect terms from a certain area of a given subject field, e.g. terms relating to punch card systems; (2) to connect terms belonging to the same facet of a given subject field, e.g. terms denoting special systems; and (3) to avoid an excessive number of terms in each class. Besides facet classifications, definitions are also included. The terms (60 per group, at most) are placed under five faceted main headings: (1) "Basic aspects of information and documentation", (2) "Documents", (3) "The activities in information and documentation", (4) "Systems in information and documentation", and (5) "Organizations and professions in information and documentation".

For the purpose of connecting this presentation with the field of education, the facet principle will be illustrated with an example from the EUDISED's (1973) Multilingual Thesaurus, which has a fully realized, albeit crude, facet classification and which, therefore, constitutes further development of the ERIC Thesaurus. The EUDISED Thesaurus is divided into 20 main facets. One of them is called "Documentation" and consists of two subfacets, namely "Information, Service" and "Index, Bibliography". Under the latter are ordered eight facets, the second of which contains ten subfacets from which "Thesaurus" is chosen as an example (Box 2).

"Thesaurus" is, among other things, related to "Vocabulary" as indicated in the alphabetically listed RT terms. There is a close correspondence between the two facets, although "Thesaurus" is more close to "Dictionary" and "Semantics" while "Vocabulary" is more related to a "Word" facet, as indicated by some two-word terms derived from "Word" ("Word frequency", "Word list"). This is an example of the possibility within the faceted structure of relating terms horisontally compared with vertical relations expressed by hierarchies (Box 1). To be hierarchically structured, Box 2 would have to contain NT relations.

Box 2. Example of a facet:
"Thesaurus" from the EUDISED classification



Thesaurus	Vocabulary
BT: Reference material	
RT: Dictionary	RT: Lexicology
Lexicology	Terminology
Semantics	Thesaurus
Terminology	Word
Vocabulary	Word frequency
	Word list

BT = Broader Term, RT = Related Term

Artandi (1970) provides a summary of research and theories in classification. She exemplifies a classification of the behavioural sciences by a proposal made by Altmann & Riessler: "Unit of study", "Dynamic-static properties of units", "Energy", "Transformation processes", "Intensity", "Distribution in time", "Distribution in space", and "Ecological setting".

The best-known example of language facets is probably Roget's Thesaurus of English Words and Phrases, which was first published more than a century ago. It is structured according to six main facets: "Abstract relations", "Space", "Intellect", "Volition", and "Affections". These in turn, are subdivided into about twenty subfacets (see Browning, 1971).

Different subject fields show different facet structure. Moreover, one and the same field may be differently faceted, depending on the classifier's view of the world. In spite of such obvious difficulties in the work on a universal classification system, the search for such a system has not ceased in library science (cf. Richmond, 1972).

2.3 Structural relations

Classification systems, whether structured in hierarchies or in facets, may be regarded as semantic in the sense that they organize concepts according to generic relationships, synonymy, antonymy, and so on. The concepts refer to relations between lexical words considered independently of context. In classification a semantic relation is usually expressed by the relationship between concepts and their properties. In this connection, therefore, it would be possible to distinguish between semantic and structural relations, where *structural* means context-dependent. A structural relation, then, is one that holds between concepts.

Analysing the content of document on the basis of its structural relations entails the advantage that the choice of keywords is restricted to a system of rules which may, to a great extent, increase the reliability between indexers in the document description process. The most obvious advantage, however, is that the document's own structuring of its content is represented. A classification is deductively imposed on a document, while analysis of structure admits an inductive generation of concepts by which their meaning is defined through their structural relations. This type of analysis is also known as "concept analysis".

One of the better-known adaptations of concept analysis in information science is the PRECIS system (PREserved Context Index System). It has been used at the British National Bibliography (BNB) since 1971 (Wellisch, 1977). During 1978 it was introduced in Swedish libraries. The different principles characterizing PRECIS (cf. Austin, 1977) will shortly be outlined below, following the main characteristics as presented in Box 3.

Box 3. Examples of structural relations: PRECIS indexing

Index string:	*Role operators:*
(0) United States	(0) Location
(1) aircraft industries	(1) Key system
(p) personnel $h unskilled	(p) Part/property
(2) training $i in-service $v by $w of	(2) Action
(3) foremen	(3) Agent
	$h non-lead direct difference
	$i lead direct difference
	$v downward reading
	$w upward reading

Entries:

Aircraft industries. United States.
Unskilled personnel. In-service training by foremen.

Personnel. Aircraft industries. United States.
Unskilled personnel. In-service training by foremen.

Training. Unskilled personnel. Aircraft industries.
United States. In-service training by foremen.

In-service training. Unskilled personnel. Aircraft
industries. United States. By foremen.

A typical PRECIS entry consists of a "heading" composed of a "lead" and its "qualifier" plus a "display". Cross references are possible in that more than one concept in an entry may become a "lead". This is done through a rotation mechanism called "shunting". The mechanism is displayed in Box 3 under "Entries". From the heading "In-service training of unskilled personnel in the American aircraft industries" have been derived four leads. The structural relationships are preserved by a numerical and alphabetical code system (The "Role operators" in Box 3). These operators define roles and links at different levels. The first level places the document in a context (e.g. geographic area) and determines to which observed system (subject field) the concept belongs (the operators 0 and 1 respectively). The second level accounts for the relevant data. This is followed by syntactic and semantic codes, which serve to localize the positions of the concepts, e.g. in a title, and which also function as keys to the original structure.

In the example in Box 3 the lead "personnel" has been coded as a part or a property of "aircraft industries". The adjective "unskilled" has a "non-lead direct difference" code ($h), because it cannot be a lead entry. "Training" is an action, which may be directly differentiated in a lead ($i). The agent "foremen" does not seem to qualify for a lead position.

The operators are used as functors, whose task is to point to the "roles" of the important concepts. The coding of those structural relations is manually performed where the $ sign is used to prevent misinterpretation during the computerized shunting procedure.

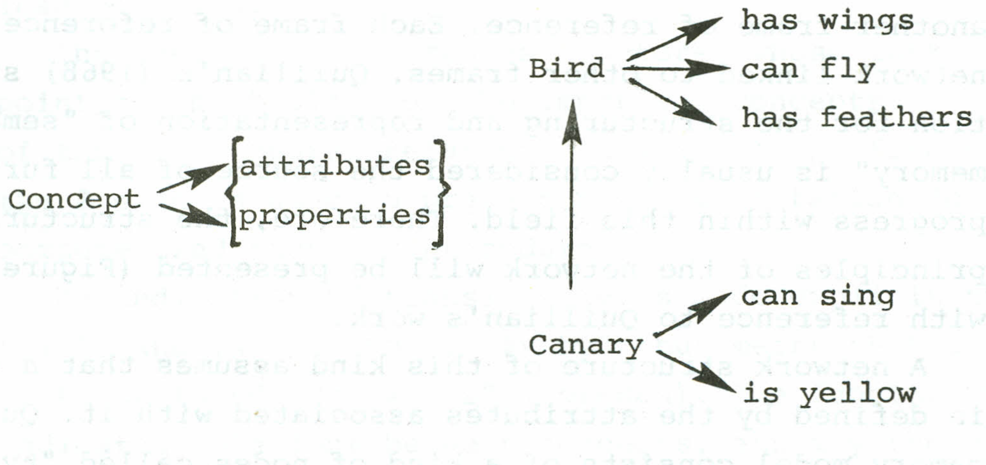
An indication of transitivity is made explicit through the $v and $w operators. "$v by" thus means that the agent is to be found "downwards" (foremen). The functor "by" indicates relations between concepts, and may therefore be regarded as an explicit operationalization of the structural relations principle. A good deal of thinking in terms of a classification scheme remains, however, probably due to the fact that the system was developed mainly for libraries (the processing of cards). Instead of "personnel" being categorized as "object" in the above example it is represented as a part or a property. The functor "of" serves as an explicit operationalization of semantic relationships. But thinking in terms of roles has contributed to document analysis in that the aims of a document, which are not always made explicit in a title, can be (manually) indexed. For example, the agent "foremen" has obviously been retrieved from some other place in the document discussed.

2.4 Networks

The structures described as *networks* in connection with information science may be regarded as systems based on associative links. With respect to the associationistic principle the networks have a psycholinguistic foundation. For the purpose of developing models for storing information (memory structure) several simulation programs have been constructed. Attempts have been made to build in a capacity for answering questions, asked from the point of view of one frame of reference, by using information from another frame of reference. Each frame of reference is a network linked to other frames. Quillian's (1968) suggestion for the structuring and representation of "semantic memory" is usually considered the source of all further progress within this field. Therefore, the structuring principles of the network will be presented (Figure 1) with reference to Quillian's work.

A network structure of this kind assumes that a concept is defined by the attributes associated with it. Quillian's memory model consists of a kind of nodes called "type nodes" (concepts) to which are assigned associative links, called "token nodes" (attributes). Thus the type nodes are names (labels) whose meaning is specified only through the attributes. Each concept-attribute frame is a net which may differ from other nets in abstraction (level). The access to concepts of higher order depends on the linkages between such frames. The associationistic principle (according to Figure 1) assumes that humans do associate "yellow" with "canary" and not with "bird" in the first place, which means that the attributes on one level do not lead to a concept on a next higher level.

The semantic network of the described type operates synthetically, from bottom to top. The associative, context-independent activation of the nodes calls for special problems to be solved by the researchers. Quillian's problem was to develop a "Teachable Language Comprehender", in which the formation of concepts relied heavily on a disambiguation process. To solve this, Quillian had to exclude the kind of semantic meaning empirically established by Osgood. Instead the meaning of the concepts refers to lexically specified semantics. Connected with the problem of lexical specification is the amount of "semantic primitives" required to include all necessary information. But as long as the network does not build on contextual relationships between the concepts, the information searcher cannot use the net for inference making.



a) basic assumption b) lexical specification

Figure 1. Example of a structuring principle in semantic networks

From Quillian (1968) it is obvious that the construction of the model of semantic memory makes use of a manual coding of certain relationships between words that humans are assumed to utilize in the understanding of the meaning of sentences. Although Quillian's (1968, p. 231) terms "subject", "object" and "modifier" do not always correspond to those of linguistics, it cannot be denied that a syntactic order is basic for the specification of lexico-semantic meaning. One of the followers of Quillian deserving mention might, therefore, be Woods, whose contribution within the domain of network structuring is a syntactic parsing system, in which the "coding" takes place automatically.

Woods' (1973) problem area is restricted to moon geology, and his LUNAR system is able to answer questions within this context to a very high degree of precision. While Quillian's network stores semantic information, Woods' network stores grammatical information. His "Augumented Transition Network" (ATN) grammar is in the first place meant to be a context-free grammar ("general expression" grammar). But the augmentation makes it capable of handling expressions of natural language, and turns it into a non context-free grammar.

The nodes symbolize the states in a hierarchic tree, whereas the links are symbolized by arcs to which are assigned arbitrary actions and conditions for the transition process. In this kind of network, the pathway between nodes is guided by a "register-setting" (set of variables) that works along with the parsing procedure. Since the transition depends on specified conditions, there has to be some kind of disambiguation even in this model. The ambiguities refer to what possible successor there is to a given configuration. Thus the registers contain lists of "syntactic primitives" which may or may not make up a high level unit. In the parsing process the "most likely" arc is decided upon. This means that the ATN grammar network operates interactively, making use of the strongest information from either syntax or semantics at each step. Since there are as many possible successor configurations as there are start states, Woods avoids the problem of defining the grammatical assumptions underlying the network.

A third type of network, strongly founded on Quillian's work, is the one represented by Schank (1972; 1973), taking conceptual relations into account. In his "Conceptual Dependency" model constructed for the purpose of understanding natural language texts, Schank specifies the relations holding between dependent and independent concepts and between conceptualizations, a new terminology for syntactic and semantic classification.

The direction of the links in this kind of network depends on what kind of relationship is assumed between the concepts. The process of defining these relationships relies on a conceptual cue system, where, e.g., "Picture Producers" and "Picture Aiders" reflect the assumption of an image-like memory structure. Although the building up of the conceptual net is made via "primitive ACTs" (which in fact are "general") indicating the need for an identification of event structures, the dependencies are based on Fillmore's (1968) cases. This makes the model static in kind, since it builds on philosophical assumptions about language, and as such it is not suitable as a model of human information processing, which requires a psychological foundation.

Schank's main contribution to research on information structuring, however, is the template-like approach, leading away from primitivization and towards models based on canonical representation. The various theoretical views of memory models are to be found primarily in the field of cognitive psychology. But most of this research can be found in the field of artificial intelligence (AI), whose main goal is to develop mechanisms for logical deduction, i.e., the application of rules of inference to statements made in a formal language, whose semantics is well specified. These restrictions, due to the restricted "intelligence" of the computer, have led to unfortunate analogies concerning theories of the structure of human memory.

2.5 Schema

By the mid 1970's the associative memory models began to be replaced by a structural approach, deriving from Bartlett's (1932) suggestion that the remembering process is schematic. In order to clarify the term "schema", some statements from Cofer's (1976) presentation of research on memory capacity will be summarized first.

The human capacity for remembering information depends on our capacity for coding it. These capabilities are most likely individual, based on the individual's strategies for processing information. The individual uses language in order to construct propositions which describe events that he has observed. The information expressed by these propositions is coded and stored in so-called long-term memory. The observed events constitute experiences, so when one talks about different "worlds of experience", this indicates the assumption that individuals have acquired different representations of propositions. Such a set of propositions may be characterized by means of a *schema.*

This hypothesis concerning propositions can be supported by several experiments on memory structure which have shown that the syntactic form (the manifest level) of a sentence does not have any crucial importance for retention (see Greene, 1977). It is the semantic relations that are retained. These relations seem to be selected and transformed according to a model characterized by a "role" or a "case" structure (the latent level). Furthermore, Kintsch (1974) has pointed out that the verb determines the extent to which sentences are confused in a so-called "recognition experiment", a result which supports Wearing's (1972) and Reid's (1974) argument that the verbs are only indirectly represented at a latent level.

A model for the analysis of latent structures has been developed and tested (Bierschenk & Bierschenk, 1976; B. Bierschenk, 1977a). It is context-oriented and based on the assumption that the manifest level can be used for the construction of a schema suited for the analysis of latent dimensions. In this model, too, the function of the verb as an organizer of concepts and conceptual relations (abstractions) is of crucial importance.

It seems appropriate to mention that a schema as a model of cognitive structure should not be confused with the notion "frame". This term refers to AI applications (Winston, 1975) of research on perception. The frame hypothesis constitutes the basis of studies of "size", "distance" and "form", phenomena that may be preprogrammed and thus imply that actions are known in advance. The schema is a structural model based on directiveness and adaptation.

A representation of information based on the schema principle is rational and saves memory space. As opposed to the explicit network structure, the schema structure is a result of abstractions, i.e. symbols and relations between symbols. Thus information can be procedurally embedded within the structure of a schema. Instead of concentrating on how information has to be re-structured at every question, a schema model attempts to find out what should be activated in order for the system to be able to give an adequate answer. The schema model eliminates the need for "activation from the bottom" and is thus based on heterhierarchical functioning, which implies utilization of cues in lower domains in order to signal the activation of a certain component of the schema, which can then be applied to the data. Moreover, the model has psychological relevance concerning both "recognition" and "recall".

The model employed to illustrate the schema principle is the Agent-action-Object model, AaO, (for a description, see Chapter 4). The three components of this model can be used for action-oriented schematic representation. If, for example, the aim is to study what agents act through given actions towards given goals, this may be schematically represented as in Figure 2.

It can be seen that there are two place holders, so-called default variables. What values these default variables will assume when the components are activated depends on what can be realized by the action "to write". For example, the question mark to the left may be replaced by "researcher" and the one to the right by "reports".

A representation of the values manifested here ("The researcher writes reports") would, e.g. from the point of view of a sentence schema, be made in terms of Subject-predicate-Object and the model would, instead of AaO, be called the SvO model, because the components of the schema are assigned syntactical terms. Another linguistic model derived from the same schematic structure would be the N_1 v N_2 model provided with semantic, predicate-argument based variables.

The characteristic feature of a schema is that its components are always present. They must, however, be provided with variables, which means that a schema is operationalized. But the single values of the variables are dependent on the context (domain) within which the schema is to be activated. Therefore, it is important to make clear the schema against which interpretations are made.

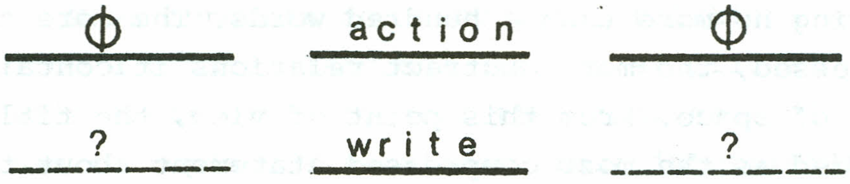


Figure 2. Example of an action schema

3. SOME MODELS OF REPRESENTATION IN THE ORGANIZATION OF INFORMATION

Information can be represented with different degrees of complexity. The representation of a given type of information can be different for different purposes, which, among other things, is indicated by the type of document in which it is presented (research reports, journal articles, handbooks, etc.). Through transformations information can be re-structured. Transformations are part of all intellectual activity, constituting a cognitive process which can manifest itself in a language structure. Thus each transformation implies a change of the language structure and, consequently, there is a connection between level of complexity and language structure. The author of a scientific report may transform the text himself to make it appropriate as a journal article, or he may write an abstract of it using no more than a hundred words. The more the text is compressed, the more abstract relations it contains per unit of space. From this point of view, the title may be regarded as the most compressed statement about the content of a work. Moreover, in an I&D system the representation of information relies heavily on the fact that someone (often called a documentalist) other than the author himself performs the transformations required for document descriptions for different purposes, e.g. writing abstracts or indexing through the assignment of some characterizing words (keywords, descriptors), often based on interpretation of the title alone. As a rule, such descriptions form the sole basis for communication between the author of the document and the information searcher. Document descriptions, therefore, are "surfaces of communication", whose language has the most central function in this kind of communication.

3.1 Intermediate languages

Language may, very broadly, be described as a means of communication. In this respect, natural and artificial languages do not differ. Their communicative capacity and function, however, differ according to the degree of formalization employed. Generally speaking, in relation to artificial languages, natural language is characterized by an abundance of variations and alternative interpretations, necessary for its function as a means of communication between human beings. A greater degree of formalization makes for greater precision and clarity, thus restricting the number of possible interpretations. An artificial language is characterized by standardization of vocabulary and rules, whose meaning must be unambiguously definable. Thus high-level languages (FORTRAN, ALGOL, LISP, etc.) display a more elaborated treatment of semantics because they are based on a strictly formalized logic, i.e. rules for manipulating statements. The direct opposite to this may be exemplified by the language used in natural discourse.

When concepts and conceptual relations are to be translated from a natural language into an artificial one, difficulties will arise owing to just that sharpness of definitions that has to replace several different interpretations possible in natural language expressions. Such transformations are processed at different levels (with more or less formal and explicitly stated changes), which does not make it easy to define the borderline where a natural language becomes artificial.

What ought to be focused on, however, is the mechanism that underlies the transition process. This mechanism forms the basis of the languages that have been developed in order to make possible access to documents, i.e. documentary languages (concerning this term, see Wersig & Neveling, 1976, p. 67). These languages are in principle as numerous as are information systems. Documentary languages, such as PRECIS (see Chapter 2.3), have a variety of structures, implying that the differences are comparable to those to be found in natural languages; in a sense, they are an abstract reflection of them. The differences in structure apply to documents (direct description) as well as to descriptions of documents (indirect description) (Coyaud, 1966, p. 127).

A language must have a lexicon and a set of rules. In spite of many attemps to formulate abstracts of varying length for I&D systems, it has not been possible to formulate rules specifying how such paraphrasing of an original document should be done. To be sure, an abstract describes a document, but the language employed cannot be termed a documentary language. Furthermore, the abstract is, besides the full text of a document, the only type of description that consists of complete sentences. The next higher lever of description would be the title, which, in general, has a reduced syntax and which could be regarded as the beginning of an artificial language. In addition, the title is the level of description most frequently drawn on in determining a document's content and in the generation of descriptive terms, both manually and automatically. Therefore, the title is the communicative interface between the document and the indexer, having a key function as the last "station" before the content is transferred to new media.

Irrespective of how many stages of abstraction are used in the transformation of document content, these descriptions may be considered variants of the language used within one and the same type of medium. Similarly, other media have their own language variants, e.g. the computer languages. The linkage between document and computer in a computerized I&D system is performed through some kind of communication, which starts at the title level and is connected to the "surface" of the computer, i.e. to a symbolic language. This language is then transformed into a machine language, which is the internal language of the computer medium. Thus communication between media is also performed through languages (with a lexicon and systems of rules), which consequently may be called *intermediate*. The intermediate function refers to a medium between languages as well as to a medium between the author of a document and the information searcher (cf. Coyaud's term "language intermédiare" in Coyaud, 1966, pp. 18-19). The structures of those kinds of language are usually represented in a thesaurus.

As mentioned in Chapter 2, the thesaurus supplies descriptors for indexing documents. The concept of *indexing* is defined in Coyaud & Siot-Decauville (1967, p. 40) as

> "la traduction de documents écrits vers leur représentation dans un langage documentaire."

Indexing is a process which starts with recognition and identification of content, followed by definition by means of a statement ("This report is about ..."). This statement is then represented by index terms, classification number, or descriptors taken from a thesaurus.

The main problem with such "informal interpretation of a document" (impressionistic content analysis) is that explicit description and operationalization of the interpretation process cannot be achieved (see Sparck Jones & Kay, 1973, p. 18). The cognitive model used by the indexer is usually unknown, even to the indexer himself (Robinson, 1977, p. 170), which means that errors made in the representation of facts cannot be controlled. Since the transformation functions have not been explicitly formulated, there is no telling in what respect a document's representation form differs from its original form.

In document description both librarians and information scientists and linguists talk about "indexing language" (see e.g. Sharp, 1967; Sparck Jones & Kay, 1973; Soergel, 1974). As mentioned above, indexing is a cognitive process, whose "language" has not been made explicit; further, it cannot without difficulties be made explicit. The control mechanisms available are the vocabulary in the subject index, the classification scheme and the thesaurus. The indexer makes use of these means in describing a document. He is the channel, while the means of control are the media.

As such they function as variants of an intermediate language. Language has here been defined as a "means of communication" having a lexicon and rules. Since it is possible to talk about more or less communicative variants of language, the communicative ability of a language may be defined through the structure that characterizes each set of lexicon (vocabulary) and rules (grammar). The structure of an intermediate language is represented through the thesaurus. The standardized terminology in the field of information and documentation (Wersig & Neveling, 1976, p. 118) defines *thesaurus* as

> "A controlled and dynamic *documentary language* containing semantically and generically related *terms*, which comprehensively covers a specific domain of knowledge."

Although a thesaurus is defined as a *language* in the passage just quoted, only the structure of the vocabulary is emphasized. Its dynamic properties, however, equally are important, especially in computer-based I&D systems, where the rules stating how the terms could and should be combined for search and retrieval have to be made explicit.

The above discussion, it is hoped, will have made it clear that a presentation of intermediate languages within I&D should involve a description of the organization and function of thesauri and other more or less structured vocabularies in relation to the role they ought to play in the communication between document and information searcher.

Since the use of a thesaurus implies active and explicit assignment of descriptors to documents, no attention will be paid to classification systems (see Chapter 2) in the subsequent presentation.

3.2 The thesaurus as a means of communication

A thesaurus, according to the above outline, will here be regarded as a language, whose purpose is to make communication possible, primarily in computer-based I&D systems. From the literature concerning the construction of information systems it is evident that there is a certain confusion as regards terminology in the field, which several authors (e.g. Fairthorne, 1969) have noted as a typical feature. Many of the statements made and positions taken are no doubt due to the fact that the field is a relatively new branch of science (having existed for no longer than approximately 20 years), and that people working within this area represent different traditions. The borderlines between the disciplines involved are vague. At the same time an integration between, e.g., library science, on the one hand, and general and computational linguistics, on the other, would be valuable. One result of this line of thought is Sparck Jones & Kay's (1973) attempt to show the extent to which linguistics and information science were actually integrated. But neither at that time (1973) nor in a more recent survey (1977) do the two authors seem to have recognized the key role played by the thesaurus as a linguistic phenomenon. They state (1973, p. 46) that the most important linguistic interests lie

> "1. in the treatment of the text of the document,
> 2. in the formulation of the description text,
> 3. in the treatment of the description text."

The linguistically interesting things that are built into the thesaurus, making possible the formulation and treatment referred to, are not discussed. There seems to be a tendency to focus on indexing as an activity in itself instead of on the medium on which it is dependent. Karlgren (1977) discusses the positions taken by Sparck Jones & Kay but, unfortunately, his position is not very precise either. For example, it is not clear if the author makes a difference between a language for description and one for retrieval. However, Karlgren makes an important point in stating that the retrieval process, as it functions today, is not treated as a linguistic problem, since retrieval is usually the result of a matching. A similar distinction between linguistic and non-linguistic means of characterizing methods for automatic document analysis is made by Coyaud & Siot-Decauville (1967). These distinctions refer to the process itself and the linguistic prerequisites have not been considered.

The following presentation is based on the statement that the success of the search and retrieval process depends on the organization of the thesaurus (cf. Chapter 1). Structural relations within and between the entries and the combinations of them in the search process must be stated in such a way that they can represent an adequate conception of the subject field. Further, it is suggested that "linguistic" and "non-linguistic" are inappropriate concepts in the present context. It would be more adequate to discuss these relationships along a continuum. This implies that varying degrees of structuring are proposed to exist in intermediate languages as well as in others. A language with a low degree of structuring is supposed to have a weak communication capacity.

The degree of structuring may be expressed in three dimensions. The first dimension concerns the *terminology*, which refers to the selection and organization of the relevant terms to be incorporated into the subject field. The *syntactic* dimension refers to the structural relations that are made explicit within and between terms. The *conceptual* dimension refers to the concepts and conceptual relations formed, which, in a title, represent the content of a document. Accordingly, *content* is defined on the basis of the conceptual model employed in the analysis of the verbal expressions under consideration (see e.g. Osgood et al., 1957; Krippendorff, 1969; B. Bierschenk, 1978a). The representation of content may be based on, e.g., the framework of the subject, the theory or the psychology of science, or on combinations of these aspects. (Cf. the discussion on schemata in Chapter 2.5.)

This model for the study of the field is conceived differently from what is usually found in the literature, and so no directly relevant references can be given. However, for a general outline of the various developmental stages within information science, the reader is referred to published volumes of the Annual Review of Information Science and Technology, in particular to the chapters that deal with "Automated Language Processing", "Content Analysis, Specification and Control", and "Document Description and Representation". Impressively analytic work concerning the general orientation in documentary languages and their structure, as compared to natural languages, has been done by Coyaud (1966). But one has to keep in mind that his work should be judged in the light of research on automatic translation, where, in a very deep sense, it represents an attempt to find universal features in documentary languages.

A general frame of reference can also be obtained from B. Bierschenk (1973; 1974a) and Salton (1971), concerning information systems and their way of functioning. Furthermore, thesauri themselves often give quite good and concrete descriptions of their usage. Some of them have been discussed here.

By and large, statistical methods will be kept out of the present discussion. They are primarily used for the estimation of significant words in attribution, and for the generation of frequency dictionaries, etc., based on abstracts or full texts.

3.3 Degree of structuring in thesauri

An organized terminology with the lowest degree of structuring may be said to be an alphabetical list of terms, selected as significant for a certain field of information. In the infancy of information science such a term index was the only control tool in indexing. Indexing problems grew in time along with growing fields of information. Expanded subject areas required expanding terminology for their description. At the same time the development of the computer-based I&D systems entailed, as a consequence, the necessity of a greater degree of formalization in the intermediate language than before. In the late 1950's and early 1960's various studies of the indexing process were performed. Among other things, they pointed to problems as regards the consistency and selection of terms (see Annual Review of Information Science and Technology, 1967, Chapter 4). In this connection a discussion started about one-word and multi-word descriptors and relations between index terms ("coordinating indexing") (see e.g. Thesaurus of ERIC Descriptors, 1975, p. XIX). In order to solve these problems, computers became tools for automatic extraction of index terms, mainly from titles of scientific documents. Thus, two techniques for the generation of terminology can be distinguished: (1) a manual technique, involving the use of controlled word lists, and (2) an automatic, uncontrolled technique. Investigations of automatic generation showed that the best descriptor was a two-word term, but also that the terminology became too bulky and unwieldy, due to, among other things, variations in the words denoting one and the same phenomenon. These experiences led to techniques for controlling the size of the vocabulary. Rules were needed which stated explicitly which terms and term combinations were important and should be allowed from the point of view of subject description, and which were possible from the point of view of language structure. By the late 1950's the so-called KWIC indexes (Key Words In Context) were created for the selection of terms. Each word in a title could become an index entry, except certain structural words, called separators. The group of words between one separator and another contained keywords and their context.

Now, when it came to creating an intermediate language which could be represented in the form of a thesaurus, it was not only the structure of the subject (see Chapter 2) or lexicographic aspects that were to be considered. Of importance were also various aspects related to the system itself. One of them was the functioning of the terminology in automatic identification of document descriptions by matching against descriptors. Another aspect was automatic indexing for content analysis (e.g. Stone et. al., 1966) of different document texts by matching against terms in natural language. It became necessary to have the terms normalized, which meant representing them in the shape of base forms. Programs for automatic suffix elimination were developed, and stop lists determined which words should not be regarded as terms. Another thing required in automatic identification was a method for standardization of universal linguistic components, able to "pick up" terms with identical morphological structure. The method is called truncation. Outputs from searches with truncated terms make it evident that it is a delicate problem to identify relevant verbal constructions, at the same time avoiding irrelevant ones (see B. Bierschenk, 1973, Chapter 7). A thesaurus that relies only on a terminology, no matter how it might be selected and organized, has a very low degree of structuring. Normalization and standardization do not solve, e.g., problems of homography or synonymy. Further, a term extracted from a KWIC index to be incorporated into a thesaurus loses its context when its lexical meaning has been determined.

The most obvious lack of communication concerning terminologically based thesauri is that the meanings of the terms are unknown. This has frustrating consequences for the information search. The matching process is only based on identity (same pattern) or partial identity (e.g. in truncation of word stems), i.e. on a coincidence of characters. Boolean algebra is used to differentiate between the existence or non-existence of terms, but structural relations are beyond its capacity.

Thus the lowest degree of structuring is to be found in automatically generated term indexes, which are not based on explicitly formulated cognitive models but may have an explicitly described syntax. The ERIC Thesaurus has tried to make use of the KWIC technique for terminological control through its "Rotated Descriptor Display", in which the allowed combinations of the descriptors included are given. But this explicit syntax does not address the problem of cognition.

Manually generated thesauri are based on implicit cognitive models, and the structuring of the terminology is performed through an implicit syntax. In order not to lose the indexers' and the subject experts' knowledge of the structural relations holding between the terms, a system of rules was introduced, relating the terms to each other (the "related term" symbol). The relations expressed concern synonymy derived from the conception of the subject field, while the implicit syntax is restricted to denote the co-existence of the terms with regard to a certain aspect of a subject field. The language structure is a noun phrase consisting of a main concept and its attribute, mostly in two-word combinations. The search logic connects such a noun phrase with another or defines an intersection which allows part of the two phrases to coincide.

An example of the utilization of an explicitly described syntax is the automatic analysis aided by phrases as presented by Hillman & Kasarda (1969), in which explicit syntactic relations in the form of fixed compounds can be extracted from documents and also be retrieved through matching of a phrase in a search query. However, a problem in automatic syntactic analysis is that many phrases relevant to document description are not matched because of different embeddings in natural language. Further, the difference between a phrase and a multi-word term is slight. The syntactic analyses performed (e.g. Salton, 1962; Klein & Simmons, 1963) utilize so-called function words to demarcate phrases. In scientific texts it can be assumed that the context between them would be of restricted length.

Therefore, it is probably a correct statement by Salton (1968) that simpler methods of analysis, i.e. such as are not syntax-based, would give at least as good results in document retrieval, since none of the methods can detect conceptual relations.

Questionable results obtained from automatic techniques for term extraction and experiences of low reliability values in manual indexing and syntactic analysis could provide the appropriate reason for the appreciation of the PRECIS system by library scientists and librarians. The system together with its applications is demonstrated in Wellisch (1977).

The general relational system in the PRECIS thesaurus contains equivalence, hierarchical relations and associative relations. Austin (1977, p. 3) claims that the system relies heavily on linguistic principles. To a large extent however, it seems to draw on library routines. Coding (indexing) is performed manually, but the "shunting" technique is computerized. This technique for producing index entries resembles that of KWIC, although the number of KWIC entries depends on the number of terms in a title. The PRECIS manual requires certain roles to be specified in the syntax-based relational system called "concept analysis". This term has been chosen because into the system has been built a dependency structure which, besides indicating notions like agent, action, etc., also defines attributive dependencies. The noun phrase as a block is kept apart from the transitivity relation that is assigned to the action term in the form of "by" or "of" ("downward reading component" and "upward reading component", respectively).

The attention paid to PRECIS indexing may be seen as an indication of a phenomenon that Sparck Jones and Kay seem to be somewhat surprised at, namely that the linguistic theories developed and tested in the late 1960's and early 1970's have exerted such limited influence on document description in information science. The libraries developed their own systems, among other things because of difficulties in integrating new methods into existing systems. By the time the ASLIB-Cranfield, the SMART and the MEDLARS projects were running, linguistic theory was central to language research and discussion. This conclusion can be drawn from surveys in the Annual Review of Information Science and Technology between 1967 and 1970. (Compare also Sharp, 1967, with Bobrow et al., 1967.) Sager (1977, p. 76) states that linguists are primarily working with theories for sentence generation, while information scientists are primarily dealing with recognition. It seems that a closer connection between linguistics and information science was not established until Fillmore (1968) appeared, probably due to the indications of cognitive features in his case theory.

The real purpose of PRECIS as a document description system is not quite clear, perhaps owing to confusing mixture of viewpoints from library science and linguistics. Undoubtedly, however, PRECIS is a step in the right direction, due to its use of syntax for coding dependency structures.

A step in the development towards a higher degree of structuring is the facet-based thesaurus, which makes it possible to connect aspects of a subject field with a certain specific relationship to each other (entities and their properties, substances and their reactions, etc.). This can be expressed both syntactically and logically. One example of an intermediate language characterized by an explicit syntax and explicit philosophical relations is SYNTOL (SYNtagmatic Organization Language), described by Coyaud (1966). A SYNTOL analysis of a text can be performed at several levels, e.g. at a morphemic and a syntagmatic level. The morphemes are analysed both analytically and synthetically. The analytic relations are made up of four "formal" classes, i.e. they are of the philosophical-logical type ("Prédicats", "Entités", "Actions", "États"). The synthetic relations relate the morphemes dynamically or statically, with the aid of syntactic conditions.

A SYNTOL syntagm constitutes a representation of a factual condition. It consists of two lexemes, whose syntactic relations to one another are explicit. An assertion paradigm ("enoncé complet") is in SYNTOL a kind of schematic representation model for documentation ("représentation documentaire") in which the "verb component" is used only to indicate the terminological relation that is to be stored. No doubt, this language employs a cognitive approach, pointing towards the kind of conceptual representation that artificial intelligence is concerned with. SYNTOL represents an attempt to create a highly abstract and general system (it was born in the spirit of the universalists), and the same theoretical idea seems to underlie Sager's LSP system (Linguistic String Parser). Although focusing on the structure of specific subject fields, her work is based on a similar paradigm (see Sager, 1977).

LSP represents an intermediate language which is highly structured from both a linguistic and a subject theoretical point of view. It is an example of the possibilities provided by computational linguistics to analyse scientific texts. LSP uses advanced linguistic techniques in combination with statistics. The analytical system has been set up empirically, with the aid of a linguistic structure. Implicit relations within the subject field are translated into the linguistic format.

Sager (1977, p. 86) writes:

> "... scientific reporting is concerned with establishing causal connections between events."

Verbs have fundamental importance in the generation of "events".

Sager uses the verbs (e.g. represented by chemical processes) for establishing the relationships between concepts (here represented by,e.g., states of chemical substances). Her response to Sparck Jones & Kay's (1973) request concerning automatic "deep structure" analysis by means of links and roles indicates interesting possibilities of development within the field of thesaurus construction. In a display (Sager, 1977, p. 90) are demonstrated clusters of "events", expressed by clusters of verbs functioning as operators together with the arguments represented in the form of the "roles" that chemical substances play in relation to each other.

It is quite easy to interprete this model as being dynamic because of its process orientation with respect to the subject field. The underlying schema should be of the AaO type (cf. Chapter 2.5). Therefore, it is confusing that the representation format chosen is the static predicate-argument model, i.e. N_1 v N_2, which cannot represent events, but only states (results of events). This circumstance is of general concern for thesaurus construction.

The generation of intermediate languages on the basis of natural language texts "of a more restricted kind" holds out hopes of a promising future, according to Sager. Linguistic problems are easier to handle, especially since the vocabulary in these texts is used unambiguously. However, Sager does not state explicitly that the assumptions behind her model apply to scientific work in general. To be sure, she makes the following statement (1977, p. 86):

> "Linguistically-based subfield formats are one answer to the question of underlying representation. While they are based on selectional constraints that operate in particular science outfields they have certain features which may be common to many science fields."

However, whether due to an inadequate model of representation or not, she does not seem to realize that the possible importance of her model for information science lies in the fact that it is aimed at the representation of "cognition", which is what information science should deal with. Without it "re-cognition" becomes unimportant. The extent to which language structures can also represent cognitive structures is an important research concern. In its attempts to approach that problem area, cognitive psychology in the latter half of the 1970's has had a valuable impact on information science (see Damerau, 1976).

4. A METHOD OF GENERATION BASED ON FUNCTIONAL RELATIONS

In the previous chapter it was suggested that an intermediate language which represents the cognitive structure in a message possesses a higher degree of precision and structuring than natural language. Such structuring should, therefore, be the goal of every system whose task is to convey abstracted information. Not until terms and their syntactic relations are grounded in an explicitly specified cognitive model can they function as units of representation in an intermediate language.

4.1 Starting-points for the construction of a model

Scientific concepts are communicated through scientific documents, whose various statements are based on empirical observations about events (Sager, 1977; B. Bierschenk, 1978b). The manifest representation of assertions about events (statements) may be described as a sentence, defined as Noun$_1$-verb-Noun$_2$, where the verb denotes the relation between the two nouns. A *sentence* will in the following discussion be referred to as N_1 v N_2. These symbols can be used in a description of how semantic relations are marked in natural language, a reason why this linguistic format has been of relevance to information science. N_1 and N_2 represent labels (see Chapter 2.4) which are used to designate "sets of information", varying in extent. Therefore, they may also be called "extensions" (see Lewis, 1972, p. 174). The way in which an extension depends on another is generally denoted by functions. In this sense, the ν symbol is a function. An extension may also be regarded as an argument which can take different values, expressed in the form of attributes. In this connection Lewis (1972, p. 177) talks about "intensions". He writes:

> "Things are name extensions and values of name intensions; sets of things are common-noun extensions and values of common-noun intensions; sequences of things are assignment coordinates of indices. Change the underlying set of things and we change the set of extensions, indices and carnapian intensions."

With Lewis's formulation as a starting-point, *intension* is defined as

> the properties connoted by a term

and *extension* as

> the class of objects designated by a specific term denotation.

Results from research on memory structure (see Cofer, 1976) seem to indicate that there exists an abstract representation of text in memory. What is stored seems to be an internal representation of a proposition. For multi-argument sentences it is easy to supply the argument(s) missing. In developing a model for the representation of a proposition, the goal should be to be able to use the context for supplying the missing parts. Thus the N_1 v N_2 paradigm should be supplemented with a default variable (ϕ), which is a place holder for missing arguments. For example, van Dijk (1977, p. 133) says that

> "... propositions may be 'present' without being (fully) expressed in the surface structure of the discourse."

This observation, formulated in connection with a "discourse model" according to Krippendorff (1969), has been applied in the ANACONDA system concerning the coding of verbal answers obtained from interviews (I. Bierschenk, 1977).

In principle, the N_1 v N_2 paradigm could be used as a model of representation even in applications of information science, since information may also be defined as being propositional (van Dijk, 1977, p. 133). But the logic implied in the N_1 v N_2 paradigm is not sufficient if the purpose is to study the interdependency between different concepts in a proposition. The linguistic categories activated by this paradigm would be word classes denoting lexical meaning, i.e., the "schema" is of the semantic-logical type. For a description of how scientific information is communicated, however, a process-oriented model is required, i.e. a proposition model denoting intentions. By *intention* is meant

> attention directed towards the goal of an action.

Thus intentionality is a basic property of directed behaviour or an action.

According to Werner & Kaplan (1963) it is the Agent-action-Object model that is used in the Indo-European languages to denote intentions. A proposition about an event, a state or conceptual relations generally consists of these components. A *proposition* may therefore be described as the AaO paradigm.

The AaO paradigm has been discussed and defined from a *psycho-linguistic* point of view in Bierschenk & Bierschenk (1976, Chapter 2). Its meaning is the following:

> *Agent* is defined as action centre or goal-seeking entity making use of various resources in order to achieve its goals. This description also includes, besides single individuals, groups, organizations and abstractions.
>
> *Action* is defined as an act performed by an agent for the purpose of achieving a goal. The act defines the meaning of the AaO paradigm.
>
> *Objective* is defined as everything that an action can be directed towards or be performed with.

The components represented by the AaO paradigm should not be confused with a case model of the Fillmore type. Fillmore's model is not basically different from other philosophical models, since it structures the world mainly in semantic-logical terms.

On the other hand, the correspondence between the N_1 v N_2 and AaO models is evident in the "function component". The *a* component is required to identify the parts of the proposition. The action denotes which object(s) or goal(s) must be present in order for a proposition to be detected. But fragments of a proposition may also be present in connection with the AaO paradigm, i.e. single values may be missing and need to be supplemented. To accomplish this, the default variable is used.

Experimental results support the hypothesis concerning the fundamental importance of the AaO paradigm as a format for representing propositions (Werner & Kaplan, 1963, p. 58). Kintsch's (1974) subjects were asked to sort sentences into categories. The sorting criterion was the relation between nouns in a sentence. The study showed that it was easiest to remember a set of nouns acting as agents. The most difficult to remember was the object role. Furthermore, Kintsch's studies seem to indicate that decomposition of complex concepts into more elementary ones did not facilitate remembering or understanding. Some kind of "lexical decomposition" (p. 249) as a psychological process for retrieval is not supported by his data.

When it has been determined what constitutes a proposition in the cognitive sense, its different manifestations in language can be examined on the basis of natural data, e.g. composition of words, variations in sentences due to the number of arguments, redundancy, etc.

As has been stressed in the previous chapters, a document can be described in many ways. But each description that goes beyond purely formal bibliographic information requires the choice of a model and thus the adoption of specific assumptions about the content of the document. A model intended to represent a proposition about the scientific work communicated in a research report cannot, contrary to the view held by Sager (1977, p. 86), be "linguistically-based", but should be related as closely as possible to the theoretical foundations regarded as adequate to the research process that the content of the document is supposed to represent. The paradigm or schema chosen as the format of representation (see Chapter 2.5) is thus determined from the basic components of the research process itself. Otherwise, the model cannot adequately represent the statement that the author makes about the research process by means of the condensed proposition in the document title. Nor can the different values assumed by the components of the model be interpreted. Sager's (1977, p. 86) statement that scientific reporting deals with the establishment of causal connections between events could be transferred to a higher level of abstraction, so that instead of concerning the structure of a subject field, it would concern the structure of research itself. Relations of a higher order would be established, making it possible to derive structural and functional aspects of several subject fields and to establish connections between them.

The single events about which the researcher communicates information have appeared in a certain contextual frame, which in turn is reported (represented) in a frame of higher order. Therefore, the title may be regarded as the proposition that represents all the others in a particular information set. Similar points are made by van Dijk (1977).

It is generally accepted that "problem", "method" and "goal" are the fundamental components in the research process (see Bunge, 1967, p. 6). The "method" component explicitly denotes the way in which the search for new information is to be done. The "Problem-method-Goal" model denotes the aim (direction) in the research process, namely a conscious steering towards or a systematic and goal-oriented search for new information. An *abstracted or "schematic" proposition* will in the following discussion be referred to as the PmG paradigm. The single components of the paradigm have been presented in B. Bierschenk (1974b) from the point of view of *research theory*. Their meaning are the following:

Problem is in this context defined as something that is reflected against a scientific background and that is to be solved by scientific means for the purpose of creating new information. In this respect the problem component has a governing function in connection with scientific activity. There is (implicit) intentionality involved, which makes it possible to compare the role of this component with that of the agent in the AaO paradigm.

Method is defined as all scientific activity performed for the purpose of showing that a problem can be solved completely, partly, or not at all. Rules concerning this activity are specified, aiming at minimizing different kinds of error sources. The rules concern, in principle, the researcher's (1) way of approaching the problem, (2) planning, and (3) instrumentation. If these are fixed, the individual researcher will act stereotypically or in a scientifically sterile way. Therefore, it is of considerable importance that problems are formalized in such a way that the formalization supports the use of adequate methods. This should be done in the form of hypotheses which can be tested against different kinds of criteria.

Goal is defined as an explicit formulation of the governing idea included in the problem component. It concerns representations of goals, levels of achievement, and anticipated solutions.

The components in this abstract proposition model should, like abstractions in general, be regarded as aggregations of concepts and conceptualizations. The values (i.e. the types of terms that an argument can take) that are assigned to each component may in themselves be of different kinds, and can be further categorized and analysed. It may be appropriate to mention that the three components are not comparable with such linguistic categories as, e.g., word classes, sentence constituents, or cases. The appearance of a "verb" is readily expected under the Method component. But in fact,a research technique (which is one of several realizations of the method) may be called "intervju" (an interview), and a problem (what a researcher tries to solve) may be referred to as "att intervjua" (to interview).

Analysing a title assigned to a scientific text from the point of view of a linguistic representation of conceptual relations (here scientific concepts) implies a study of the result of an abstract representation of the text as manifested in a title. Such a study is based on cues provided by the manifest structure of the title.

Since the overt organization of the title may be fragmentary and restricted with respect to the PmG paradigm, a default variable is required even in this case, serving as a place holder for missing arguments.

In order to analyse the relations between the single components in the title, a way of indicating the component's roles and structural connections is needed, referring to the theoretical starting-points of the model.

4.2 Presentation of a recognition model

Starting on the assumption that a title is an abstracted proposition about a research process, it can be stated that the single components in the process must be distinguishable for the title to be properly understood. This understanding is likely to be strongly dependent on the structure that can be inferred between concepts representing the conceptualization. This structural relationship also has to be accessible through cues that can be discerned and organized in such a way that information processing can take place.

The basic linguistic elements used to relate concepts and to structure reality are the prepositions. Depending on the model of language under discussion, the prepositions are labelled "morphemes" or "lexemes", where the former applies to syntactic and the latter to semantic function. In order to study cognitive development and language acquisition both functions have been employed in psycholinguistics and cognitive psychology (see Brown, 1973). These kinds of study refer mainly to the structural function. The organizational function is emphasized in computer applications of text analysis, especially in information and documentation.

To give an overview of the various theories, viewpoints and applications involving prepositions would result in a "combinatorial explosion". Instead some typical approaches will be presented here, which could be of special relevance for the understanding of the development and operationalization of the present model.

As was mentioned in Chapter 3.3, the KWIC indexing technique was created for selection of terms suitable for thesaurus entries. Especially in titles these were highly relevant words or groups of words that could be automatically selected by the help of structural words called separators (Coyaud, 1966). These separators were mostly prepositions with no other role than to be a stop word for dissecting segments of various length. As in most manually performed indexing the prepositions used in this way were "dropped" when a segment had been dissected, which explains why keywords are unrelated.

Another typical approach within information science is to use prepositions (as well as other so-called "syntactic joints") as statistic variables in automatic estimation of distance between significant terms in a document vocabulary (O'Connor, 1973). It is found that precision in retrieval is surprisingly high in document passages belonging to "soft sciences", which makes the author believe in his "syntactic closeness" estimation as an even more valuable approach in the "hard sciences".

An example of the use of prepositions as separators is a program developed in order to automatically separate titles in a multilingual document base (I. Bierschenk, 1978). The study shows that a relatively small set of separators can organize to a high degree of probability an extensive data base with respect to the language of the titles. A special problem in this connection, though, is the identity existing at the expression level, as in "in", commonly used in Swedish, English and German titles. The phenomenon has been termed interlingual homography (see Allén, 1970, Introduction).

The works of Salton (1962), Klein & Simmons (1963) and Hillman & Kasarda (1969) exemplify a more explicit syntax-based use of prepositions where they serve as syntactic function words, i.e., fixed compounds can be extracted. Here the prepositions have an organizationally specified role as entries to or constituents within phrases. It is generally assumed that the access to documents is performed with a higher precision by using phrases than just keywords when searching through abstracts or extracts of document descriptions. The reason is that a phrase provides a less ambiguous interpretation of its constituents. For example, a term "program", phrased "program for graduate studies" should not be mixed up with the same term used in the context "program in computer science".

By the name of concordance the KWIC technique is widely used as an aid in solving ambiguous interpretation of "sequencies of characters" in natural language. In document retrieval, as well as in other fields dealing with the construction of vocabularies of various textual domains, such detection of phrases can evoke new insights into language structure, for example, the way Allén (1976; 1977) discusses the phenomenon of extending the borders of idioms to collocations called constructions. The "constructional tendency" of prepositions, e.g., could help to define meaning through usage. Further, insights into the algorithmic fixation in natural language should be of general importance for text processing (see also Gross, 1980).

In a document retrieval context the prepositions are nothing but signals to certain organizational features of text. The "structural" and "functional" properties of the prepositions refer to textual structure and the function of pointing to certain constructions, as in many other applications of automatic text processing. An attempt to use prepositions as pointers to "concepts" (the structural relations approach) is shown in the PRECIS system (cf. Chapter 2.3). But the relations referred to are a mixture between syntactic roles (coded, e.g., through "by") and generic-semantic roles (e.g. through "of"), i.e. between two linguistic models. It is also evident that the so-called "Key system" or "Location" refers to a geographical area, specified by the help of a "knowledge of the world" strategy in which prepositions like "in", "at" and "from" are often employed. The classification of those entries is not made out of the language structure.

In syntactic analysis, especially when automatically performed as in so-called parsing, one of the most delicate problems is the ambiguity inherent in prepositions. According to Woods (1973) this holds particularly in the modifying prepositional phrases, such as "I saw the man in the park with the telescope". Following Woods' discussion the with-phrase might modify "I", "the man" or "the park" and thus it is ambiguous, although, syntactically, it modifies the last noun phrase. Woods' ATN grammar contains lexical information specifying what verbs can take what kinds of arguments. Since "see" may have "optical instruments" among its arguments, the solution of this ambiguity problem is found through semantic rules in a predicate-argument model. Woods' parser is an example of the mutual relationship existing between syntax and semantics in the interpretation of natural language.

In a computational treatment of case grammar Friedman (1973) distinguishes between "real" and "meaning-bearing" prepositions and others. The first kind are selected from the lexicon. The others are either predictable from verb (as in "laugh at") and case or are inserted by transformations ("of", "for", "by"). The choice of preposition for a particular case can be stated as part of the case frame feature. Friedman obviously differentiates between prepositions as verb particles and as case pointers. The same differentiation is made in Helbig & Schenkel (1973), whose work is aimed at constructing a dictionary displaying the "Valenz" (values or case frames) of German verbs, for which a case-like model is particularly suitable. The authors propose that the particle governs the object as in "Er wartet auf den Freund" (He is waiting for the friend), whereas the preposition, as in "Er wartet auf dem Bahnhof" (He is waiting at the station), governs the adverbial phrase. There is an important difference, though, between these two approaches. While Helbig and Schenkel state that they refer to syntactic values, Friedman's model relies on semantic feature values ("lexicalist-case grammar").

A semantically based usage of prepositions in information science context is an experiment reported in Braun & Schwind (1975). An attempt is made to extract index phrases in which the relations between concepts are stored in a hierarchical semantic network. The phrases dealt with are noun phrases (nouns, adjectives and adverbs). The prepositions are used as pointers to supplements of noun or adjective terms. This network principle is based on logic, from a set-theoretical viewpoint, in which the PREP notation is handled like a variable whose values are names of prepositional relations, e.g. OF or IN. The network has been constructed on the basis of a certain textual domain and tested by thesaurus matching, including simple morphological analysis. An overall result is that this logic-based semantics was not sufficient to correctly extract phrases of the sought kind. Information about the succession of terms has to be incorporated in the rules, a so-called "syntactic filtering".

A recent view taken by AI researchers is that semantic relations are procedurally defined. A simulated dialogue with a robot by name of SHRDLU invented by Winograd (1972) was the starting point. The robot shows a full "understanding" of a very restricted world of blocks by moving them to different positions in a "knowledge space" defined through computer topology. In the analysis of processes, as, e.g., descriptions of games (chess, cards etc.), it is of importance to be able to ascertain that a task has been correctly understood, which is done through truth-value checks of states and state changes. Thus these should be a logical pattern in natural language enabling the logical conception to arise. The prepositions play a key role in this procedure. Thereby two kinds of prepositions can be distinguished, namely those which refer to states ("in", "on" etc.) and those referring to procedures (such as "from" and "to"). A Swedish example in this direction is Cedvall's (1977) semantic analysis of computer accessibility to instructions for card playing made in natural language.

Human cognition is supposed to be based on spatially organized representations of phenomena, as pointed out by, among others, Piaget & Inhelder (1956) and Miller & Johnson-Laird (1976). Basic experiments have been performed by Piaget & Inhelder, indicating that the first spatial understanding in children is topological. Children obviously understand proximal properties like order, demarcation and continuity. Not until later developmental stages have been reached do properties such as angles, parallelism and distance become comprehensible. This theoretical view has been tested by means of the acquisition of prepositions (Grimm, 1975). The developmental sequence of spatial prepositions concurs with the spatio-geometric properties shown by Piaget.

Piaget's (1963) opinion is that adults build up implicit cognitive representations (schemata) consisting of coordinates. These schemata are employed for orientation in space and time. Thus it can be hypothesized that prepositions constitute the functions both in the process of building up a cognitive schema and in the producing of it. Based on this view, a scientific title can be regarded as a manifest representation of a "scientific cognition" which by means of prepositions can be re-cognized.

A typical title of a scientific text can be:

En analys av titlar (1)
(An analysis of titles)

The "scientific event" underlying this title may be described in terms of statements like

Jag analyserar titlar	(I analyse titles)
Jag har analyserat titlar	(I have analysed titles)

The "event" condensed here derives both the agent and the action from *En analys* (An analysis). The transformational level is marked by the preposition *av* (of). Before the transformation the verb form indicated the concept *titlar* (titles) as being an "object". This role is still to be discerned after the transformation through the function of the preposition *av* (of).

The goal of scientific inquiry, however, is not to handle persons or solid objects, but to deal with problems. However, problems also imply that there are possible solutions, which means that problems determine the research process in the same way as the object in a sentence determines what type of verb may be selected. Therefore, it might be justified to have the label "object" replaced by the label "problem". As previously mentioned, Problems include intentions; consequently, the role of the component in the PmG paradigm is comparable with the Agent in the AaO paradigm. The preposition *av* (of) then functions as an operator for the Problem component. When the problem has been identified, *En analys* (An analysis) remains to be analysed. This part of the title can now be given an unambiguous interpretation, i.e. it denotes the scientific event, as manifested in the methods or means used. These two components can be fitted into the PmG paradigm as follows:

Kategori (Category)	PROBLEM (PROBLEM)	METOD (METHOD)	MÅL (GOAL)
Operator (Operator)	av (of)		
Representation (Representation)	titlar (titles)	En analys (An analysis)	ϕ

Scientific activity cannot be equated with a determinable object or determinable problems, but should be defined as a strategy, i.e. a way of tackling problems. The development of new methods and instruments increases the individual researcher's possibilities of creating new information. For this reason, the informative value of the title increases if it contains information on the research strategy or technique used. If the information is expanded, so that the analytical technique is made explicit, the title can be given as

En analys av titlar med en kodningsalgoritm (2)
(An analysis of titles with a coding algorithm)

As shown sofar, the research strategy is *En analys* (An analysis). The preposition *med* (with), however, specifies in more detail what plans, techniques, or instruments have been employed. The new information can be fitted into the PmG paradigm as follows:

Kategori (Category)	PROBLEM (PROBLEM)	METOD (METHOD)	(INSTRUMENT (INSTRUMENT))	MÅL (GOAL)
Operator (Operator)	av (of)		med (with)	
Representation (Representation)	titlar (titles)	En analys (An analysis)	en kodningsalgoritm (a coding algorithm)	ϕ

The brackets around the Instrument component denote that it is optional. But means or instruments play a central role in science. It is therefore reasonable to assume that *med en kodningsalgoritm* (with a coding algorithm) is an explicit expression of what researcher X does, i.e. his way of analysing *titlar* (titles), and that this is conceived to be an important piece of information to communicate. Therefore, all concepts denoting means may be arranged under the Instrument component. In research these are seldom solid objects (tools). However, they do have a more concrete function in connection with the method. This is the reason, in the example under discussion, why the Instrument component is made explicit and placed between Method and Goal. (The goal determines the instrumentation of a method.)

If the title is further expanded, so that it also contains an explicit statement of the goal, it may appear in the following form:

En analys av titlar med en kodningsalgoritm för begreppsigenkänning (3)
(An analysis of titles with a coding algorithm for concept recognition)

The goal is, in this case, to recognize concepts. As example (3) shows, the preposition *för* (for) denotes this intention, i.e. it gives the reason why a certain act that requires certain instruments has been performed. The information about the goal can be fitted into the PmG paradigm as follows:

Kategori (Category)	PROBLEM (PROBLEM)	METOD (METHOD)	(INSTRUMENT (INSTRUMENT))	MÅL (GOAL)
Operator (Operator)	av (of)		med (with)	för (for)
Representation	titlar	En analys	en kodningsalgoritm	begreppsigenkänning
(Representation)	(titles)	(An analysis)	(a coding algorithm)	(concept recognition)

This title is a representation of the schema when totally filled with values together with an optional component. In "reality", one or more of the positions (arguments) will be filled with defaults. For example, the optional component may not appear when the others are activated.

What has been described so far are the main components of the PmG paradigm, illustrated by a particular research strategy or schema. The representation of this strategy is the result of a course of events, which should be seen as movement in space and time. But since every kind of movement requires a description in space and time, this would be trivial information in a title. Therefore, in general, the concrete place and time of a particular research activity are not specified (and hardly ever is information given concerning the place and time of the writing of the report itself). If, nevertheless, place and time are indicated, the most appropriate thing, from a linguistic point of view, would be to let space and time become determiners to the sentence itself, i.e. to the verb component. But in connection with the transformation of a "concrete" natural language expression of a course of events, e.g. "Jag har analyserat titlar i flera månader" (I have analysed titles for several months), into the abstracted intermediate form "En analys av titlar" (An analysis of titles), the temporal aspect of the verb form and the denotation of time by the prepositional phrase are nullified. Transferred to the PmG paradigm, space and time would determine the method. However, since the method itself determines the results of the research, space and time are irrelevant concepts.

Solid objects are considered with respect to their relative location (see Ralph, 1977). Depending on the dimension in which the object is demarcated, this may be expressed with prepositions, for example *i* (in) for space and *på* (at) for surfaces. The child learns to see relations between solid objects and to express these relations by means of prepositions. The ability to form concepts (abstractions) comes later, although the prepositions used to relate abstractions are the same. For example, Grimm (1975) has shown that there are developmental stages when the word field and the concept field do not correspond (e.g., in the conception of the concrete-local and abstract-temporal dimensions) resulting in a substitution of prepositions until a stabilization occurs. Language conventions are often the reason why a preposition with a plainly two-dimensional function, e.g. *på* (at), is used to denote a more abstract relationship. Prepositions denoting space are also used to specify time, since time, too, has direction and extent. In different kinds of automatic analysis of natural language, as has been discussed above, this ambiguous use of prepositions is a great disadvantage for the determination of the meaning of the component that follows. But since scientific titles convey abstract relations on an intermediate level, the ambiguity that is necessary in more concrete contexts is eliminated in the same way as certain aspects of verbs are no longer relevant after a certain abstraction has been performed. This is further discussed in the following section.

Research focuses on problems which are multi-faceted. This implies that a scientific title gives expression to multidimensional phenomena. But problems, too, may be determined as being part of a problem area which incorporates a time dimension. (For a discussion, see Miller &

Johnson-Laird, 1976.) This is expressed in the title:

En analys av begrepp i titlar från fyra decennier (4)
(An analysis of concepts in titles from
four decades)

The preposition *i* (in) determines where the *begrepp* (concepts) are to be found, i.e. in titles and in no other type of text. The preposition *i* (in) here denotes a demarcation in that it specifies the contextual domain of the problem under consideration. The *titlar* (titles) must not be regarded as a concrete place or container in this conception, which is why a philosophical model of inclusion (cf. Ralph, 1977) will not do. The preposition *från* (from) is used in contexts of space, denoting a starting-point. It has the same meaning when used on the time dimension, i.e. it demarcates the temporal range. These relations may be incorporated into the PmG paradigm as follows:

Kategori (Category)	PROBLEM (PROBLEM)			METOD (METHOD)	MÅL (GOAL)
Operator (Operator)	av (of)	i (in)	från (from)		
Representation (Representation)	begrepp (concepts)			En analys (An analysis)	ϕ
		titlar (titles)			
			fyra decennier (four decades)		

As can be seen from the above examples, a strict sequential order is described. The last demarcation specifies *begrepp* (concepts) in such a way that it concerns not only titles but also a particular period of time. There seems to be a need for a time dimension in order to discriminate between experiences from different periods of time.

It should also be mentioned that geographic places, too, are regarded as abstract concepts in this model. Depending on where in the schema such a name is inserted, it demarcates the component under consideration. In the title "Cognitive Science in Sweden" the geographic name specifies the problem area dealt with and should not be conceived as Locality. Locality refers to a model based on concrete actions which must have a "place of performance". The model employed here assumes that the processing of scientific concepts does not need the concrete information. Instead "in Sweden" denotes a demarcation incorporating both space and time and thus serving as a help for the information searcher in specifying the conception of "Cognitive Science" contextually. (In fact, the scientific information conveyed is the conception of that subject at that time.) This approach is discussed in greater detail in the next section.

According to Oller & Sales (1969) the principle of concentric order seems to be of general relevance in the analysis of sentences, in the sense that the most specific information is located farthest away from the sentence kernel. This principle will here be employed in the sense that the important organizational function of the prepositions will form the basis for automatic demarcation and determination of format. The operationalization will be demonstrated in the following section.

4.3 Operationalization of the model

The creation of order presupposes a schema within which the order is to be set up. Thus the task of creating order within titles first of all implies the determination of what constitutes a title. The function of a title has been discussed previously and it may seem superfluous to go further into the concept of *title* at this stage. But in the coding and editing of thousands of titles written in several languages, it soon becomes evident that titles may be structured in many ways, perhaps as a consequence of their function. One can distinguish the length of the title, the form and colour of the letters, the title's localization on the cover of the document, typological variations concerning capitals and small letters in main titles and subtitles, different kinds of punctuation marks to organize the various entities in the title, and so on. But in writing instructions for automatic organization of the entities all dimensions in the title are not considered, because many of them have no significance in the representation of document content. However, when large bodies of data are to be handled, a unified format must be determined which indicates the limits within which interpretation and inference are allowed.

The abstractions and the relations between them conveyed by a scientific title are the visual result of the author's conceptualization at a certain point in time. The processes that preceded this conceptualization, i.e. the formation of the scientific concepts, are no longer distinguishable, at least not in one and the same title. Thus a reconstruction of the processes involved is hardly possible once the conceptualization has been completed (cf. Kintsch, 1974).

A title represents one or more conceptualizations, which in their manifest form can consist of more or less complex language structures. A title consisting of one word consequently expresses a conceptualization whose structural relations are implicit, while a highly structured title explicitly indicates such relations, e.g. through prepositions. What is important for the development of a system of rules is to mark a detectable boundary for a conceptualization. In order to avoid difficulties of interpretation it will be necessary to utilize purely orthographic marks. In the instructions concerning the punching of the material it was said that a subtitle should be separated from the main title by a full stop (which is usually avoided in titles due to considerations of layout). Semicolon, colon and dash also serve as demarcators. Titles such as the following express two conceptualizations:

Frågor kring studiedagar - en enkätundersökning (5)
(Questions about study days - a questionnaire
investigation)

De akademiska undervisningsformerna (6)
Universitetspedagogik
(The academic teaching forms)
(University pedagogy)

The titles are authentic. In order to maintain the structure of the Swedish wording, it has been necessary to give a word-by-word translation. This applies also to the examples given in Chapter 6.

Another (albeit rarely seen) type of boundary marker is represented by conjunctions coordinating two conceptualizations, i.e. they function as connectors.

En empirisk studie av kognitiv utveckling samt (7)
(An empirical study of cognitive development and [also]
en kritisk analys av intelligensbegreppet
a critical analysis of the intelligence concept)

This boundary marker merely separates two conceptualizations from one another rather than connecting them. The difference between these and the ones that are demarcated by a punctuation mark is that they are explicitly marked as coordinated. But the boundary marker *samt* (and [also]) signals that a new conceptualization has to be coded; consequently, *samt* (and [also]) is regarded as a disconnector.

Conjunctions functioning as "real" conjunctions coordinate concepts within the same conceptualization. Commas, added by the punching, also function as connectors:

Inlärningsmaskiner och programmerade hjälpmedel (8)
(Learning machines and aids for programmed
instruction)

Familj, skola, samhälle (9)
(Family, school, society)

Mätning av språkfärdighet i (10)
(The measurement of language proficiency in
engelska och tyska
English and German)

These instructions and decisions belong to the kind of order-creating rules that may be called *demarcation rules*.

Another form of demarcation which serves an editing function is the use of brackets. In order not to disconnect such an entity from the one it belongs to (in the function of explaining, etc.), it became necessary to disregard certain prepositions within those entities. By this arrangement they are automatically assigned to the nearest preceding entity.

The demarcation rules have provided the outer framework of the model. Two types of rules then form the basis for the conceptual coding, namely *stop rules* and *structuring rules*.

As indicated in the previous section, the analytical method employed rests on certain basic assumptions concerning the function of the prepositions in a title, which is to indicate, or point forward towards, certain types of concepts. In this function their position relative to one another is of considerable importance. From the model it is obvious that the prepositions ("pointers") belonging to the main components, are essentially *av* (of), *med* (with), and *för* (for). These prepositions propel "the action" forward, thus expressing the transitivity in the paradigm ("the horizontal level"). Prepositions demarcating the main components are located between them and in a specified order. As a consequence, the concepts they point forward towards are ordered "vertically" under the nearest preceding main preposition. This state of affairs can be compared with, e.g., Abelson's (1973) and Faughts's (1977) models of "belief systems". Faught (1977, p. 5) writes:

> "Human use intensional constructs such as beliefs and intentions to order the environment and direct their behavior."

Intentions are realized through actions and are constructed through conceptualizations. In the PmG paradigm, this corresponds to the components being distinguished by means of *av* (of), *med* (with), and *för* (for), on the one hand, and being demarcated and defined by means of prepositions such as *i* (in), *på* (at) and *från* (from), on the other.

Within the theoretical context presented in the previous section, the former type of prepositions will be called *intentional*. On the assumption that a concept is perceived as an extension or denotation with respect to the properties and characteristics that form the intensions or connections of the concept, an explicit specification of the concept's intension is regarded as an extension. This extension should be considered to be spatial and as such it functions as a demarcation, i.e. as a visually signalled extent. Therefore, the latter type of prepositions will be called *extensional*.

This is illustrated in the figure below:

INTENTION	av (of)	med (with)	för (for)
EXTENSION	i (in) på (at) från (from) . . .	i (in) på (at) från (from) . . .	i (in) på (at) från (from) . . .

Figure 3. Sketch of the organizing principles in titles

In Chapter 4.2 the use of the so-called locative prepositions in the localization of abstractions was mentioned. With respect to the original meaning of these prepositions it was pointed out that their ambiguity when used with objects of a concrete kind is nullified when the objects themselves are of an abstract kind. The usefulness of the assumption of prepositions as functions, which is interesting from the point of view of computational linguistics, has a psychological relevance in this model. A function does not have any meaning of its own. As soon as it has performed its task of relating two concepts (comparable to placing something in a system of coordinates, see Chapter 4.2), it loses its importance. Only the functional relation between the concepts remains.

A prepositional framework suitable for the representation of "inner cases" (logical or cognitive) in natural Swedish has been discussed and denied by Brodda (1973).

The "problematic" (ambiguous) prepositions correspond to those that in the present context are termed intentional. This natural-language conception of the prepositions has great impact on the interpretation of them when used in different contexts, why a closer look at the distribution and usage of them may be of interest here.

In Nusvensk Frekvensordbok (Frequency Dictionary of Present-Day Swedish), part 4 (NFO 4), covering the language of the daily press (see Allén, et al., 1980), the following principal facets can be extracted with respect to the intentional prepositions (as they are called here). Phrases beginning with *av* (of) are mainly said to indicate "origin in time and space", "source of event" and "focused object". Phrases beginning with *med* (with) are chiefly explained as having "associative function". In the function "taking place through" (instrumental) they seem to have a relatively low frequency. The main functions of *för* (for) are given as "with indicated purpose" and "with influence on someone".

These examples provide a good illustration for the discussion in Chapter 3 concerning differences between natural and artificial language. The variability in natural-language expressions allows for variations in interpretation, which can be utilized for more or less finely graded analyses. In this case, determining the "meaning" of a preposition depends on its context. The variation that can be shown with such an analysis, however, concerns the domain of usage. The more artificial the text under analysis is, the greater the precision required in the definitions of the prepositions. In this perspective such variations in "meaning" as those presented in NFO 4 can be seen as paraphrases, which are stabilized at the intermediate level, i.e. a basic function emerges. Thus the reduced syntactic structure in titles must not give rise to differences in interpretation. This principle has been employed at an even more abstract level, the logical, in order to successfully translate between natural English and French. In Wilks' (1973) templatebased model prepositions are transformed to a few canonical representations (cf. Schank's ACTs) covering the meaning of several stereotypes.

It follows that an interesting objective of research would be the extent to which the same basic function is assumed and used in different texts. The preposition *av* (of), according to NFO 4, may be compared with the function *av* (of) in scientific titles. The "natural" texts focus on a kind of agentive meaning which, as already mentioned, does not obtain in titles. Instead, an Object function emerges, called "problem". The function *för* (for) evidently denotes intention, i.e. a representation of a goal, and also a Result or a Recipient aspect, both of which may be included in Goal. The meaning "with influence on someone" might concern persons as the goal of an action. The preposition *med* (with), which in NFO 4 primarily denotes an associative function, has in the PmG paradigm been defined as denoting an instrument. Differences may be caused by different interpretations of the term "Instrument", depending on the model within which the term belongs.

With a kind of semantic model, the two titles "Märkning av metalliska material med radioaktiva spårämnen" (Marking of metallic materials with radioactive trace elements) and "Radioaktiva spårämnen för märkning av metalliska material" (Radioactive trace elements for marking of metallic materials) are homeosemic according to Karlgren (1974), because their underlying structure is the same. With this model of interpretation the instrument in the first title is case-based, i.e., meaning is lexically specified as to what can be means in this process. This subject-oriented meaning is then transferred in the interpretation of the second with no explicit use of the preposition.

Another example (personal communication with Dr. B. Brodda) of the interpretation of instruments will be given starting from the title "Svetsning av stål med hög kolhalt" (Welding of steel with a high carbon content). Based on a philosophical-linguistic model the with-phrase would be interpreted as having an associative function, i.e. the "kolhalt" (carbon content) is a property of the "stål" (steel). Underlying this interpretation is a N_1 v N_2 model which considers "has-a links" as static. However, seen with the PmG view the "kolhalt" (carbon content) is an instrument to the "svetsning" (welding) process. Conveyed as scientific information about industrial methods and means this title expresses its author's conception of the property of the steel being instrumental in the process: The welding is probably performed in a certain specific way when the steel has this property. Thus the model governs the interpretation.

A syntactic model of interpreting is employed by Welin's (1974) survey of titles with the perspective of correctly relating prepositional phrases by the parsing in which an explicit use of the prepositions is required. In this study, too, the "problematic" prepositions are considered, although with a syntactic structural approach. An example is a title like "Intresse bland lärare för engelska" (Interest among teachers in English). At least in Swedish this title is structurally ambiguous since the preposition "för" (in) is inherent in "intresse" (interest) thus being defined by the verbal noun's constructural property. In this case the preposition is regarded as a morphemic explication of the main word, and thus its relative position to it raises the difficulty in recognizing the structure.

Ambiguity seems to be a relative concept. Structural ambiguity may, in addition, refer to semantic as well as syntactic models of interpretation. For information processing such models may be inadequate, if the conception of prepositions is not adapted to an explicitly formulated information (cognition) oriented model of analysis, that is, a schema. Such a model may contain an unambiguous usage of the organizational structure of titles as well.

In the present analysis the "schematic" organization in titles for automatic coding is emphasized (see Fig. 3, p. 68). The sequential order among the prepositions is employed in this theoretical context by way of numerical codes expressing the relations between the "axes" in a coordinate system as follows:

En analys av begrepp i titlar (11)
(An analysis of concepts in titles)
40 30 33

The concepts expressed by the intentional prepositions are assigned numerical code numbers ending with "0". The associated extensional concept is assigned another number, where the sequential order begins with a number other than zero. The following concept would have the numerical code number 34, and so on. The Method component is assigned code number 40. This system is set up in such a way that the sequential order is algorithmically fixed. For practical reasons, the coding system has been taken over from the ANACONDA system (Bierschenk & Bierschenk, 1976, p. 40). It should be kept in mind that the ANACONDA model is applied to natural language (interviews), thus including more components than does the PmG model. Certain codes, therefore, are left empty in PmG, others are processing a generalized meaning, because of the higher level of abstraction in the PmG model. The numeric codes for conceptual coding in this study are the following:

Problem	30	and Extensions 33, 34 etc.
Method	40	
Goal	70	and Extensions 73, 74 etc.
Instrument	80	and Extensions 83, 84 etc.

The prepositions may in these titles have varying contexts, which must be precisely definable, i.e. the "length" of the prepositions must be specified. For there are a number of multi-word expressions that may be regarded as prepositions (cf., Welin, 1974, p. 138). Such a compound preposition may consist of several strings of characters (see Box 4, p. 75). Prepositions may also be part of fixed phrases, where they do not have any pointer function. Such strings have been listed in a dictionary.

(A multi-word phrase may here be preceded by an operating preposition, e.g. *inom ramen för* (within the frame of), where *ramen för* (the frame of) is specified as a multi-word phrase; in this role it does not allow *för* (of) to operate. The operating preposition is *inom* (within).)

The number and the type of fixed combinations probably differ somewhat within different subject fields. The dictionary employed here is empirically set up, i.e. specified on the basis of characteristics of the authentic material. This approach is governed by the model implying that all possible cases of prepositional constructions would be superfluous.

The doubt expressed by Welin (1974) regarding structural ambiguity (dissimilarity) is to a great extent due to the non-homogeneity in his material. Thus one may ask what the ambiguity refers to (any title from a certain period? any title from a certain subject field?). The titles examined in the present analysis have been collected from works written by a randomly selected sample of researchers from a population in which a specific definition of "researcher" has determined what titles were to be included in the analysis. A scientific title, then, represents work carried out by a researcher. Since researchers hold different posts and have specialized in different fields, differences with respect to a certain researcher and among researchers may be reflected in the titles of their different works. This may give rise to the utilization of phrases and combinations typical of more concrete expressions than would be expected in titles.

Rules for treating such different cases of combinations involving prepositions are here called stop rules. Prepositions may be combined with other prepositions, either directly (as in "Tysk skola *av i* dag" (The German school of today), "Varför lärarna inte kan vara *med i* Hem-och-skola" (Why teachers cannot join the Home-School Society)), or linked by means of a conjunction ("*För* och *mot* den nya skolan" (For and against the new school [system])). Preposition and conjunction may also form a pair ("Bakgrund *till och* tolkning av..." (Background to and interpretation of ...), or "... på lågstadiet *och på* fritidshem" (... in primary school and in centres for children's leisure activities)). Specification of such cases prevents coding errors. Otherwise the rules for concept coding will interfere with each other when concepts are linked by way of a connector (the conjunctions *och* (and), *eller* (or) or comma). In the same way as an interconnection takes place in ordinary sentence analysis, the concepts on both sides of a connector should be assigned the same numerical code number.

Before the structuring rules are presented, the rules that the dictionary operates with are given in Box 4. In the English translation of the Swedish prepositions only one alternative is given, considered the most common in the contexts in question. It should be observed that the multi-word prepositions together represent only two functions, *i* (in) and *med* (with). In the same way the main preposition *av* (of) is also represented by four other variants.

It is assumed that the pointing and ordering functions of the prepositions are dependent on the order among them. The pointing forwards, as implied by the intention, consequently means that the method governing the "movement" is placed at the very beginning, "pushing" the other parts in front of it. The first focus of the method is the problem, if *av* (of) is the first preposition, and so on. Consequently, when all intentional components have been defined, the remainder is always coded as being the method. This basic principle is demonstrated in the following coded examples:

Psykologiska analyser av militära befattningar (12)
(Psychological analyses of military posts)
40 30

Mätningar med projektiva test (13)
(Measurements with projective tests)
40 80

Läshjälp för synsvaga (14)
(Reading aid for the visually handicapped)
40 70

Box 4. Dictionary in automatic organization of concepts in titles

Prepositions are: *

i	(in)	som	(as)	mot	(towards)
av	(of)	mellan	(among)	enligt	(according to)
för	(for)	från	(from)	genom	(through)
på	(at)	rörande	(concerning)	inför	(at)
till	(to)	under	(under)	åt	(to)
med	(with)	hos	(in)	ur	(from)
om	(on)	kring	(around)	över	(on)
inom	(within)	bland	(among)	angående	(concerning)
vid	(by)	efter	(after)		

Multi-word phrases are:

typer av	(types of)
slag av	(kind of)
grad av	(degree of)
former av	(forms of)
ramen för	(the frame of)
samvariation med	(intercorrelation with)
samverkan med	(cooperation with)
ett perspektiv av	(a perspective of)
synpunkter på	(views on)
exempel på	(examples of)
redogörelse för	(account of)
aspekter på	(aspects of)

Conjunctions are:

och	(and)
eller	(or)
samt	(and [also])
resp	(respectively)
,	,

Main prepositions are:		*Function*
av	(of)	= av
om	(on)	= av
angående	(concerning)	= av
rörande	(concerning)	= av
över	(on)	= av
med	(with)	
för	(for)	

Multi-word prepositions are: *

i anslutning till	(in connection with)	= i
i samband med	(in connection with)	= i
i relation till	(in relation to)	= i
i fråga om	(concerning)	= i
med hjälp av	(by means of)	= med
med speciellt avseende på	(with special reference to)	= i
med särskild hänsyn till	(with special reference to)	= i
med särskild anknytning till	(with special relation to)	= i
med särskild inriktning på	(with special focus on)	= i
med tonvikt lagd vid	(with emphasis on)	= i
med tonvikt lagd på	(with emphasis on)	= i
kombinerad med	(combined with)	= i
jämförd med	(compared with)	= i

* Operational definition

The concepts that in these titles have been arranged under the Method component are examples of activities that have been started in order to solve a problem. Strategies, procedures, single events, etc., have been consolidated.

Many titles do not have a Method component, namely those without a preposition. The same goes for titles with an initial preposition:

Differentieringsfrågan (15)
(The differentiality problem)
30

Om kunskap (16)
(On knowledge)
30

When the title does not express an intention, i.e. when the structural relations are implicit, what is referred to is only that a certain problem area is dealt with in some way. The preposition *om* (on) is not governed by a "pushing" method and has no function in the model. For similar reasons, a terminal preposition (although very rare) has no function from this point of view:

Vad funderar barn . på? (17)
(What think children about?)

This title just expresses that a problem is dealt with without any scientific specification of how or why.

Since the system is based on distinguishing main concepts from subconcepts, the rules must also be constructed in such a way that titles of considerable length will be assigned a correct coding. A title of a certain length may include several instances of one and the same preposition. Regardless of what status the prepositions may have in the model, only one main component of the same kind can be activated. This can be illustrated by means of the following example:

Utvärdering	av	försök	med	en	variant	av	(18)
(Evaluation	of	experiments	with	a	variant	of	
40		30	80			83	

årskurslös undervisning
nongraded teaching)

The last instance of *av* (of) determines only the variables of the instrument.

Below are presented, in a variant of natural language, the algorithm for automatic coding of the concepts in the titles. The program is written in ASCII FORTRAN.

Rule 1. Control multi-word prepositions

Rule 2. Control multi-word phrases

Rule 3. Preposition and conjunction within ( ) do not operate

Rule 4. A preposition as first word does not operate

Rule 5. If a conjunction connects two prepositions, then the conjunction and the second preposition do not operate

Rule 6. If a preposition is followed by a conjunction,then the preposition does not operate

Rule 7. If a conjunction is followed by a preposition,then the preposition does not operate

Rule 8. If a preposition is followed by a preposition, then the first preposition does not operate

Rule 9. If *med* (with) or *för* (for) are repeated, then the second instance becomes subordinated (extensional)

Rule 10. Only one of *av* (of), *om* (on), *rörande* (concerning), *angående* (concerning), or *över* (on) can be the main (intentional) preposition in a clause, and must not be preceded by an extensional preposition. A "clause" is demarcated by a conjunction or a comma

Rule 11. A preposition as terminal word does not operate

Rule 12. Main rule for the initial part of a clause

40	av om rörande angående över	(of) (on) (concerning) (concerning) (on)	30
40	för	(for)	70
40	med	(with)	80
30	not main preposition		33

Explanation: This rule states that if the first preposition is intentional (*av ...*, *för*, *med*) the following element is assigned the codes 30, 70 and 80 respectively, and the element preceding the preposition is assigned code 40. If the first preposition is not intentional, then the preceding element is assigned code 30, and the following is assigned an extensional code.

Rule 13. Main rule for the non-initial part of a clause

——	av . . .	(of) . . .	30	not main preposition	33	not main preposition	34
——	för	(for)	70	not main preposition	73	not main preposition	74
——	med	(with)	80	not main preposition	83	not main preposition	84
——				not main preposition	33	av (of)	34

Explanation: When the initial part of a clause has been identified, the coding of the non-initial part is dependent on it. Thus a second preposition is extensional if it is preceded by an element governed by an intentional preposition. The following prepositions are then assigned extensional codes according to sequential order.

Rule 14: A conjunction connects elements
1

Rule 14: A conjunction connects expressions of the same type
2

It should be emphasized here that these rules have been developed for the testing of the model. A more detailed description of them together with the computer program can be found in a separate publication (Bierschenk, Bierschenk & Sternerup-Hansson, 1979). In the account of the result of the empirical test that will follow in Chapter 6, the way in which the algorithmic analysis can recognize concepts will be made clear, and outcomes, especially from the more complicated rules,will be discussed.

5. STRUCTURES IN A REPRESENTATIVE DATA BASE

The testing of an analytical method of the kind presented here cannot be performed without a data base. Depending on the purpose of the analysis, it can be organized in different ways and created by means of different techniques. The data base to which the method under discussion has been applied differs in some respects from the bases on which analyses in information science are usually based.

The data base is experimental. It contains (1) all bibliographic data concerning scientific documents produced by a sample of researchers during a period of 40 years, (2) all references cited by these researchers in the documents from this period, and (3) a linkage between the references given in the documents and an extensive interview material (4000 typed pages) concerning the researchers' grant-supported activities. A more detailed account of the data base is given below.

5.1 Organization of the data base

The goal of developing a method for the generation of an intermediate language, with the capacity to convey information within a certain field of application (here research within education), requires a decision regarding who or what should represent the field. Based on results presented in B. Bierschenk (1974), the definition of the researcher population employed here includes psychologists, educationists and sociologists. After "researcher" had been defined, a random sample was drawn from the resulting population. The knowledge represented by these researchers

(B. Bierschenk, 1979) can be retrieved in non-fugitive form from their written works (cf. Chapter 1), which have all been collected, starting from their first scientific product (Ph. L. or Ph. D. thesis). Thus this sample of works may be regarded as representative of the focus of attention in research of relevance to Swedish education.

As a rule, each work includes a so-called reference list of the various sources and publications drawn on. The references listed may give a certain idea about the kind of research information that has been of relevance. It is possible to study and control the central vocabulary in the references given. This constitutes a kind of content analysis method used in document description (see Taulbee, 1968).

An experimental data base has been set up, containing bibliographic descriptions of the works and their respective references, connected through identification codes. The descriptions are built up according to one of the international standards (American Psychological Association, 1965). To allow the processing of single pieces of information, they have been divided into several fields. The different types of information are presented in Box 5.

Box 5. Representation of bibliographic information for computational processing

Field	*Information*
1	Name of author with initials for first name and information about author function, e.g. Ed.
2	Title and subtitle of a work
3	Place of publishing
4	Publishing company
5	Year, volume, number, page
6	Name of journal, series, mimeograph
7	Other characteristics of work

Each bibliographic reference can be unambiguously identified. The identification code specifies author, sequential order of works, sequential order of references (when given), field entry number, and sequential order within field (when more than 80 columns have been used).

For an analysis of the kind for which this model was developed, among the bibliographic characteristics it is only field 2 (titles) that is of any concern. Field 7 contains information about number of pages, language used, document type, and the like. A detailed description of the works, the coding and control is given in B. Bierschenk (1979, Chapter 2 and Appendix). But some numerical data that may be of special interest will be given here.

The number of works stored in the base is 949, of which 660 (69,5 %) are written in Swedish. The English works number 249 (26,2 %), the German works 26 (2,7 %), works written in other languages being represented by 14 titles (1,4 %).

The number of references cited is 23,141. Here, English is the dominant language, represented by 12,345 (53,4 %) titles, followed by Swedish, 8,865 (38,3 %). German references total 1,153 (5 %), the figure for titles in other languages being 778 (3,3 %).

In order to create manageable and uniform sets of data and formats for several kinds of studies, the document descriptions in the work base and the reference base have been sorted according to the language in which they are written. Field 7 provides such information about the works, and a program has been developed for automatic grouping of references according to language in the reference base (I. Bierschenk, 1978). For testing the automatic coding of information in titles, the Swedish work base and the Swedish reference base were used.

Since, for an investigation of titles, it may be of interest to examine whether there is any correlation between document types and titles, some variables relevant to such a description are presented below. Those variables for which values exist are listed in Box 6.

Box 6. Types of document represented in research on education

- Research report
- Article in research journal
- Monograph
- Chapter in book, edited by someone else
- Mimeograph
- Textbook
- Article in daily or professional press
- Preface
- Symposium publication
- Bibliography
- Official Governmental Report
- Read paper, invited address

The document types "research report" and "mimeograph" can be distinguished in that reports refer mainly to project reports, published in reviewed series, printed in various departments, whereas mimeographs refer to the kind of "grey reports" that does not have this formal status.

5.2 Patterns in titles

The coding rules presented in Chapter 4 were constructed after tests had been performed on the Swedish titles of the work base. After control of some initial tests, the demarcation rules were specified. For the construction of the structuring rules the demarcations had to be edited in such a way that a full stop was surrounded by blanks. Continued testing then showed where other changes were needed. A character is not supposed to carry more than one function, which required editing discriminating between dashes and hyphens. A dash is regarded as having a demarcating function and so it had to be surrounded by blanks. Compare the following variation in coding:

```
    En studie av kreativitetsutvecklingen inom             (19)
   (A  study  of creativity development    within
    40        30                           33

    årskurserna 4-9
    the grades   4-9)   alternatively... 4 - 9

                                         30
```

In order to prevent the number 9 from being coded as a sentence of its own (which is the consequence when a concept is single), a hyphen had to be inserted.

Dashes also had to be exchanged for commas as in the following example:

```
    Lärares  erfarenheter  från förskola   - lågstadium   (20)
   (Teachers' experiences  from pre-school - primary
    30                     33                30

                 - fritidshem
    school       - centres for children's leisure activities)
                   30
```

This title is "incorrectly" constructed, and so the demarcation mark was exchanged for a character denoting connectivity. Thus, the coding should instead be:

```
    Lärares    erfarenheter  från förskola,    lågstadium   (21)
   (Teachers'  experiences   from pre-school,  primary
    30                       33                33

                fritidshem
    school,     centres for children's leisure activities)
                33
```

Editing among non-alphabetic characters also concerned the colon, whose function is that of a demarcator. When denoting the genitive, as in "*SIA:s* organisation" (SIA's organization), it was deleted, the genitive marker *s* thus being directly connected to the word stem instead. Further, a comma was replaced by an apostrophe when functioning as a decimal point ("betyget *2,3*" (the grade 2,3) was changed into ".. *2´3*"). Certain controls and editings were performed (interactively) via a terminal connecting the Department of Educational and Psychological Research in Malmö to the Computer Centre in Lund through UNIVAC's CTS (Conversational Time Sharing System). Editing involving systematic changes, e.g. moving punctuation marks, was carried out automatically.

After these controls and editings the rules were tested once again, and then the phrase dictionaries and the stop rules were specified.

The controlling steps now concern the ability of the rules to operate correctly on the material. The content in the repsective codes will be discussed in connection with their relevance to a thesaurus (Chapter 6).

In the following the patterns emerging from the titles will be presented. According to the analytical model the conceptualizations may be more or less explicitly stated. In the most explicit case they are represented by the components 30 + 40 + (80) + 70, together with possible attributes. Thus a pattern is a structural representation of a conceptualization (see the demarcation rules), which implies that certain patterns are possible, common, uncommon, or impossible. Furthermore, there is an in-built restriction in the system, among other things due to the sequential order of the subordinate codes.

For a quantitative description of patterns, "profiles" were printed containing all the existing types together with frequency counts. They showed that there are 85 different patterns, 42 of which (50 %) are unique. The latter types are not very suitable for a quantitative description of the material. Instead, the focus of interest is on certain recurrent patterns, so as to make it possible to discover regularities, which is a prerequisite for the development of algorithms for automatic analyses.

Therefore, it was decided that the reference base should be used as a control base. The work base is to a certain extent a subgroup of the reference base, and so the pattern profiles were also counted on the reference base. The references represent 241 patterns, 89 of which are unique. The number of different patterns in the references is higher than in the work base. At the same time, however, the reference patterns are characterized by a larger number of common features than the works, since only a third of them are unique. A comparison between the patterns in the two bases thus makes it possible to estimate, with some degree of confidence, the consistency of the general features of the work profiles.

In order to determine the commonality of patterns in the works and the references, a lower limit was needed. The works have been produced by 40 researchers. If the same pattern occurs either four times in the works of one person or, conversely, once in each of four persons' works, this means a frequency of 10 % in the sample. Frequencies under 10 % may be regarded as random variation. Therefore, it was decided that only patterns with a frequency of 5 or more should be considered in the comparison.

The patterns of the works were ordered according to their ranks and then compared with the ranks of the references for the respective pattern. Spearman's rank correlation was calculated and found to be high (r_S .89). (The inspection of the material made it evident that editing of the references would have resulted in an even higher correlation.) By means of criterion 5 the work patterns resulted in 19 distinct places of rank. Table 1 shows the result of this comparison.

Table 1. Patterns in titles: comparison of ranks in work base and reference base

Patterns		Rank Order		Patterns		Rank Order	
		W	R			W	R
1.	30	1	1	11.	40 80	11	11
2.	30 33	2	3	12.	40 30 30	12	13
3.	30 30	3	2	13.	40 30 33 33	13.5	18.5
4.	40 30	4	4	14.	40 30 33 34	13.5	12
5.	40 30 33	5	6	15.	30 33 33 34	15	23.5
6.	40 70	6	5	16.	30 30 33 33	16.5	16.5
7.	30 33 34	7	10	17.	30 33 33 33	16.5	15
8.	30 30 33	8	8	18.	30 33 34 34	18	16.5
9.	30 30 30	9.5	7	19.	40 70 70	19	20
10.	30 33 33	9.5	9				

W = Works, R = References

From Table 1 it is readily seen that to a great extent regular patterns exist. Moreover, the result is based on some 9,500 titles, quite a high number in this kind of study. The first six patterns may be considered the most typical and the most predictable. Variations are marginal. The first difference between the bases is the pattern at work rank 7, the second at rank 13.5, and the third at 15.

The general result of the comparison is that differences in works and references may be apparent when there are more than one extension. It does not seem to be of importance whether the Method component is activated or not. These patterns represent such specific titles as can be found in research reports and mimeographs (cf. Box 7 in the following section). That such works are not cited as often as books may, among other things, be due to the fact that they are not as accessible as are books. The pattern can be exemplified with

Rättstavningsförmågans struktur hos pojkar (22)
(Correct spelling ability structure in boys
30 33

och flickor i årskurs 4
and girls in grade 4)
33 34

In view of the covariation of patterns within the reference base and the high correlation between the bases, only the title patterns in the work base will be studied further. The analysis, therefore, is focused on the patterns themselves in the context of the coding rules, whereby Table 1 serves as the background information.

Problems in the form of single concepts of the type shown in examples (15) and (16) in Chapter 4.3 constitute the most frequent pattern, followed by a combination of two Problem concepts or a Problem together with one extension. The more complex the patterns are, the less often they appear. A pattern simultaneously activating all components in the analytical model does not seem to exist in the material. Only two intentional components are activated in one and the same title, in the first place Method + Problem, in the second place Method + Goal, and in the third place Method + Instrument. Further, when activated, Method is always activated first. Problem is the only other component that can also be activated initially. These observations lead on to the question of regularities within the patterns, regardless of frequency. For a study of the activation patterns of the components, a matrix was set up showing the sequential order of codes as seen throughout pattern types in the work titles (I. Bierschenk, 1980, p. 66). This matrix is now summarized and visualized by means of the graph in Figure 4. For this description the proportions were calculated, a lower limit for the calculation being set at a raw frequency of 10. It was further decided that only such proportions should be considered at the interpretation as were equal to or higher than .10. Therefore, the proportions do not sum up to 1.00 in the graph. As a consequence, it is easy to distinguish the general features of the patterns.

The horizontally related nodes in Figure 4 indicate the intention of the analytical paradigm, whereas the vertically related ones indicate extensions of the concepts. The

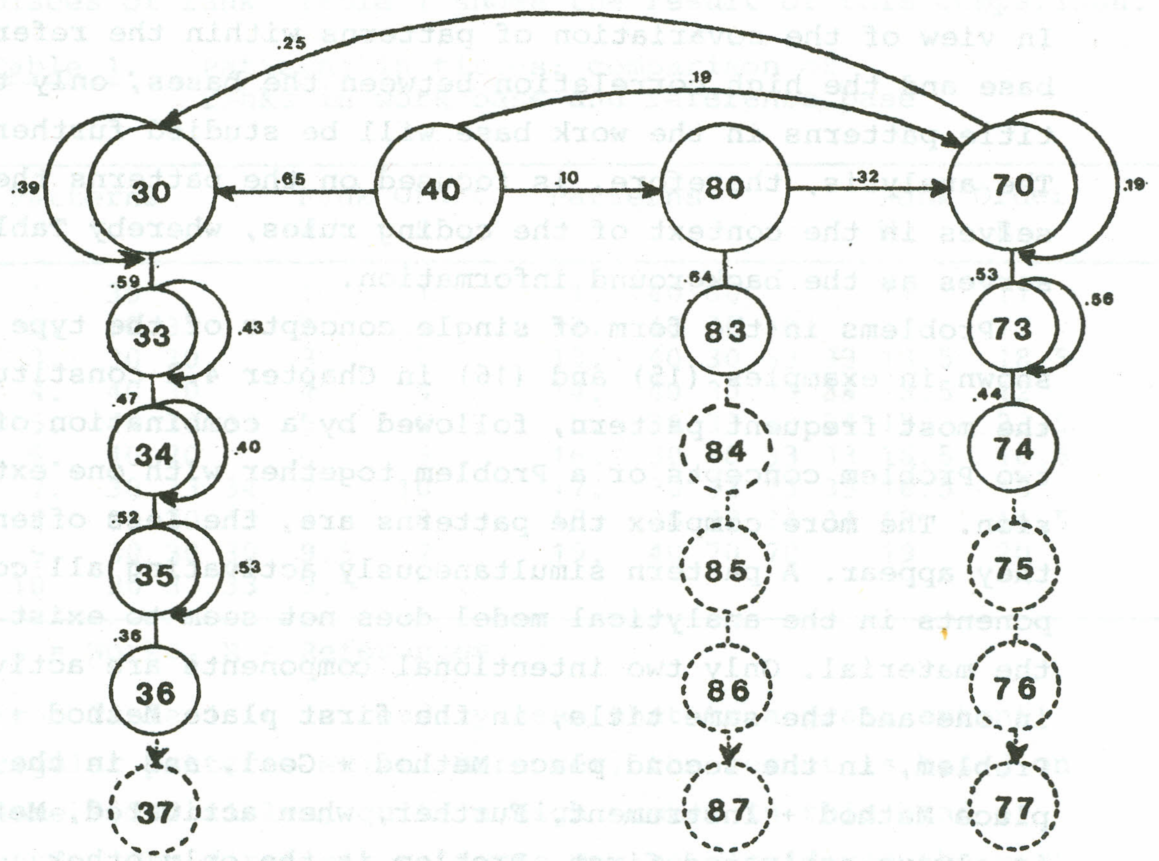


Figure 4. Graph description of titles

broken-line nodes mark the place of possible arguments according to the model. The figures at the transition indicate the proportion with which a certain argument, when appearing, is followed by the next one. The arrow from a node back to itself indicates the reflexive functions of the model.

Of all the possible patterns not many have been activated. The pattern with the highest proportion is represented by the link between Method (40) and Problem (30). There are also relatively high proportions within almost the entire Problem complex. ("Complex" is here defined as "consisting of interconnected parts".) It then emerges that Goal (70) is demarcated to a greater extent than Instrument (80), but that the tendency is higher that Instrument, when appearing, has an extension, although only one. Connectivity, however, does not appear at Instrument. Further, Instrument "governs" Goal to a greater extent than Method "governs" Instrument or Goal.

A more detailed inspection of the graph shows that only 30 and 40 appear initially. The concentration in the left nodes indicates the dominance of the Problem component. It is also apparent that the Method component is followed only by a main component, mostly 30. The main components are more frequently followed by an extension than by other main components. When Goal is followed by Problem, it may be assumed that a coordinate construction is involved, i.e. the Problem component initiates the following conceptualization, or that a concept with a double function is present (see example (49), Chapter 6). An extensional node is frequently repeated or followed by the following node in the sequential order; the lower its sequential number, the more often is this the case. This should imply that the more extensions a concept has, the more likely is it that a new main component does not follow.

Finally it should be noted that Method does not appear in final position in the titles; there is no link leading to the 40-node. This pattern is a logical consequence of the fact that Method is not preceded by a preposition, in its turn reconstructing the logical order of the research process.

As argued in Chapter 4.3, titles should reflect the activity of their authors. According to B. Bierschenk (1977b), educational researchers express a desire for methodological intensification, but also prefer working with general problems. In other words, the problem orientation is obvious. The development of new methods requires in many respects greater work intensity and time investment, which grant-supported research does not allow owing to the time limits imposed. This is probably also the reason why Instrument and Goal are not very often explicitly mentioned. Demarcating and defining problems in such a way that they become "researchable" require so much of the project time that other activities in the research process are often suppressed. This seems to be borne out by the patterns in the titles communicated by the researchers themselves.

5.3 Degree of differentiation as a function of document type

On the basis of the results accounted for in the preceding section, it appears natural to take a closer look at the works produced by the researchers, i.e. an attempt will be made to determine whether there are particular structures characterizing titles of specific types of documents. Since this type of research covers a wide range of activities, it should be expected that there is an interrelation between the construction of titles and the form of representation chosen.

In Chapter 5.1 it was mentioned that the bibliographic information has been supplemented with non-bibliographic material, including, among other things, type of document. The types that are represented in the material and which will be used here have been presented in Box 6.

As a basis for the comparison, type of pattern according to Table 1 (Chapter 5.2) is used. The same limit has been chosen, i.e. the pattern must have a frequency of at least

5 throughout all works in the Swedish language. But in order to prevent the matrix from becoming too open, such structural relations can be employed as have crystallized from the graph. Thus a grouping should be performed.

The first criterion for grouping concerns the two main patterns, namely the difference between Problem (30) and Method (40). In this way patterns with and without 40 are distinguished. Then the patterns are analysed according to intentionality or extensionality, indicating structural complexity at different levels. The 40 type is process-oriented, the 30 type problem-oriented, which means that the former relates phenomena whereas the latter describes and demarcates one and the same phenomenon. Further differentiation then results first in the groups 40 + 30, 40 + 70, and 40 + 80, representing explicit intentional relations between concepts. The corresponding pattern of the other type is a single 30, since the problem orientation is characterized by implicit intention. Thus explicitly stated intentionality forms one main group, and implicit intentionality forms the other. The degree of complexity is not assumed to increase with connective relations. This implies that combinations such as 40 + 30 + 30 and 30 + 30 + 30 within the respective groups are allowed (see the variants in Table 1, Chapter 5.2).

A further distinction is now necessary. It concerns the presence of extensions within each main group. Among the patterns in the first group it can be seen that only the 40 + 30 type is followed by extensions. If the connectivity rule is to be followed (as it should be), the pattern 40 + 30 + 33 is also formed. As a consequence the pattern 30 + 33 is given within the other main group.

The degree of complexity grows according to the concentric principle. Thus one more pattern in each main group can be formed, namely 40 + 30 + 33 + 34 and 30 + 33 + 34, respectively.

A comparison between document types will now be made with these groups as a starting-point. Eight groups could be discerned. They were ranked according to the frequency

Table 2. Proportions of pattern types for single document types

Document Type	Rank of Pattern Type *							
	1	2	3	4	5	6	7	8
1. Bibliography	.01	.01	.02					
2. Preface	.02	.02	.03			.03		
3. Chapter in book	.08	.10	.13	.04	.06	.06		
4. Journal article	.14	.15		.07	.20	.11	.15	.11
5. Professional press	.04	.03			.03	.03	.08	
6. Monograph	.25	.14	.13	.04		.34	.38	.22
7. Research report	.35	.40	.53	.71	.49	.17	.31	.67
8. Textbook	.06	.05	.02			.20		
9. Read paper	.00	.01			.03			
10. Symposium publication	.01	.02						
11. Official Governm. Report	.02	.01	.02		.11			
12. Mimeograph	.03	.05	.12	.18	.09	.06	.08	
Total Sum of Frequencies	468	206	60	45	35	35	13	9

*1. 30
2. 30 + 33
3. 40 + 30
4. 40 + 30 + 33
5. 30 + 33 + 34
6. 40 + 70
7. 40 + 80
8. 40 + 30 + 33 + 34

of the entire group. The result is shown in Table 2. The Table presents a classification of document types and a ranking order of pattern types. If each document is regarded as a sample from the work base, it may be of interest to study which pattern is the most typical for the respective samples. In order to find these patterns it is appropriate to standardize the frequencies with which the patterns appear. Such a standardization should be performed for each pattern, i.e. across the categories. The first pattern has thus been estimated by setting the raw frequency in proportion to the total sum of frequencies. Thus the proportion .006 ≈ .01 is estimated by setting the original frequency 3 to 468 ... 15 to 468. The last figure .032 is represented in the table as .03. The method for determining the typical pattern is as follows:

First the highest proportion for a certain pattern is looked up across the rows. Then it is determined for which row the pattern is the most typical. For example, the pattern 30 + 33 + 34 is typical of the categories 4 and 11. But since .20 is a higher value than .11, it can be concluded that for category 4 this pattern is the most typical.

This criterion gives the following general features. The pattern type *30* cannot be said to be typical of any particular category. The fact that titles can display more than one pattern (cf. Chapter 4.3) may be the reason why this pattern type does not differentiate between categories. The type *30 + 33* is very frequent, too. A Problem component, single or together with one extension at most, appears in all the document types and so does not function as a typical pattern representing a certain sample of documents.

When it comes to *40 + 30*, it can be considered a typical pattern. It is the most typical pattern in titles which are headings of chapters in a book edited by someone other than the author of the title. These are often books containing scientific papers. Titles of books produced by the 40 authors, each title having been produced by a single author, are characterized by other typical patterns.

The pattern *40 + 80* is typical of monographs, whereas textbooks are of the *40 + 70* type. The monographs in this material include some theses whose basic feature is the explicit statement of the instruments and techniques used for the testing of a method. The textbooks investigated state explicitly for whom or what they are intended, i.e. the goal may be certain groups of persons, a grade in school, etc.

Titles of "book types" typically contain a Method component + one more intentional component, which seems to vary systematically with document type.

The *40 + 30 + 33* pattern is typical of research reports. Thus reports have, as opposed to the others, both intention and extension explicitly stated in their titles. There are also reports with a higher degree of extensionality (the pattern 40 + 30 + 33 + 34). This, however, is not the most typical situation according to the method of determination employed here. A higher degree of extensionality is in general expressed in the titles of scientific journal articles, but the type pattern is *30 + 33 + 34*; thus there is no explicit statement of intentionality.

Further, the proportions show that there are pattern similarities between research reports and mimeographs, between journal articles and Official Government Reports, and between monographs and textbooks. But according to the method used for the determination of typical patterns, only five types are distinguished, differing from each other in so-called structural complexity. First, a group expressing explicit intentionality is determined through Method + one more main component. Second, a type pattern is determined having implicit intentionality and second-degree extensionality. Third, a further type pattern is found, characterized by explicit intentionality and first-degree extensionality. The pattern structure thus emerging is summarized in Box 7.

Box 7. Structural variation in titles: complexity characterizing type of document

Group	*Pattern Type*	*Document Type*
1	40 + 70	Textbook
	40 + 80	Monograph
	40 + 30	Chapter in book
2	30 + 33 + 34	Journal article
3	40 + 30 + 33	Research report

The structural variation in the titles, as resulting from the analysis performed, seems to have a differential effect with respect to the form of representation chosen. However, without detailed studies and experiments it is difficult to determine the extent to which the conceptual structures in the single groups correspond to the information they are intended to communicate.

In the following chapter results from the automatic coding of the titles will be presented. Examples of various difficulties will also be given.

6. CONCEPTS IN FUNCTIONALLY RELATED REGISTERS

In this chapter it will be demonstrated how the coding mechanism has worked out on authentic material. First the results of the operation strategy of the rules will be exemplified. Then the coding will be examined within the context of language structures. Finally, the concepts and their function in the registers that are to be generated will be presented and discussed. (For a definition of "register" in this context, see Chapter 1.)

The presentation in the first section will be structured according to the groups that were distinguished and described in Chapter 5.3. But not all examples per group coincide with the document type of which the pattern is typical. Titles from other document types will also be included in the presentation in order to illustrate the underlying conceptual variation of the PmG model. As has been discussed in Chapter 4, a conceptual schema which does not require detailed analyses of different elements in the language is adopted here. It is important to keep this characteristic of the PmG paradigm in mind when studying and interpreting the results of the analysis.

6.1 Coding of functional relations

A significant result of the pattern structure analysis is that most of the structures are of the types 30 and 30 + 33, including connective structures, mainly the 30 + 30 variant. A single 30 pattern has been exemplified in Chapter 4 (examples (15) - (17)) and will not be repeated here. The principles of the restricted coding, i.e. coding by means of paired numbers denoting super- and subordination within the concept complex, were also explained in Chapter 4. (24) below may serve as an illustration of a computer output with that type of pattern, as compared to (23) in which the concepts function connectively.

Psykologin och samhället (23)
(Psychology and society)

30 30

Tonårsskola i utbildningssamhälle (24)
(Teenager school in educational society)

30 33

That (23) aims at discussing the relationship between psychology and society is certainly indisputable, and as long as the author of the title does not supply any further information the two phenomena should be treated together and thus be coded as two concepts of the same order. It is also likely that such a title may entail a discussion of society from a psychological perspective, i.e. the title "Samhället och psykologin" (Society and psychology) could be as relevant as the original one. "Psykologin i relation till samhället" (Psychology in relation to society), however, would more clearly indicate the intended focus. In example (24) the extension (code 33) denotes the scope (context) within which the problem area *tonårsskola* (teenager school) is discussed. Since none of the titles within these patterns cause any trouble and further demarcations connectively marked by *och* (and) or comma do not change the result within these patterns, they will not be considered any further here. Instead, the discussion will concentrate on the pattern groups described in the previous section (see Box 7).

The first group consists of a pattern type including Method and Goal (40 + 70), Instrument (40 + 80), or Problem (40 + 30), thus representing titles with explicit intentionality. Structurally these three are similar, so there is no ranking order here. The order of exemplification has been chosen according to the degree of explicitness in the model.

Studieteknik för vuxna (25)
(Study techniques for adults)
40 70

Mål för lärarutbildning (26)
(Goals for teacher training)
40 70

Några program för elektronisk databehandling (27)
(Some programs for electronic data processing)
40 70

A typical concern in this population of researchers is to develop teaching materials and guides of several kinds. In general, they address themselves to explicitly stated goal populations (25). *Vuxna* (adults) is thus the group towards which the methodological work is directed. Another common activity is to produce goal descriptions (26) for various teaching and training purposes. The concept *mål* (goals) as a representative of a method may seem strange. But *målbeskrivning* (goal description), i.e. the activities involved in *att beskriva mål* (describing goals, to describe goals) has a clearly methodological meaning within what is called educational technology. Here, the goal is to provide *lärarutbildning* (teacher training) with a set of operationalized goals. (27) is an example of a methodological activity with a more research-oriented goal.

Some kind of instrument is often involved in the activity, even if not explicitly stated. Some such coding results are given below.

Effektivare träning med videobandspelare (28)
(More effective training with videotape recorder)
40 80

Försök med två olika typer av ordlistor (29)
(Experiments with two different types of word lists)
40 80

Familjeterapi med alkoholskadade föräldrar (30)
(Family therapy with alcoholic parents)
40 80

These examples show different kinds of concepts in an instrumental function. Technical aids are typical instruments in education and teaching (28). Educational research has during the last 15 years been characterized as classroom experiments. As a model for testing and evaluation it has often used prototypes of teaching materials (29). Both instructional and educational instruments, as well as research instruments, are therefore to be considered means of attaining goals which in this context are implicit.

Another method is representative of researchers involved in therapy (30). In this example a group of persons functions as instrument. It can be noted that the title "Familjeterpai *för* alkoholskadade föräldrar" (Family therapy *for* alcoholic parents) would have implied that the persons themselves had been the objective focused on (i.e. the goal population). In example (30) the instrumental function indicates that the method as such (and its possible effects) is in focus. *Alkoholskadade föräldrar* (alcoholic parents) has in the development and discussion of the therapeutic method the function of means. In empirical research the variables of analysis form the instrument itself. A title like "Integrering av barn med handikapp" (Integration of children with handicaps) is thus an example of a *med* (with) phrase coded as being instrumental, and not associative (cf. the discussion in Chapter 4.3).

Finally, some examples from the third variant in this pattern group are given.

Mätning av mental prestationsutveckling (31)
(Measurement of mental performance development)

40 30

Reflexioner om vardagsinlärning (32)
(Reflections on everyday learning)

40 30

Grunddragen av den svenska militära undervisningens (33)
(An outline of the Swedish military education

40 30

historia
history)

Several kinds of measurement are connected with the more experimentally oriented part of this population of researchers (31). *Mental prestationsutveckling* (mental performance development) is presumed to be a well-defined problem, since it can be measured. What is meant by *vardagsinlärning* (everyday learning) does not yet seem to lead anywhere further than to reflections. Here the choice of preposition is important. It is even more important in example (33), and there is no reason to believe that this author is not conscious of the implications of his use of preposition. On the contrary. If, namely, the author had used *i* (in), the pattern would have been 30 + 33, and *grunddragen* (outline) would have been coded as problem. With knowledge of this particular author's field of activity it can be said that the method is to *dra ut* (lay bare) *grunderna* (the outlines) of the history of Swedish education. The concept *grunddragen* (outline) is a Swedish example of how events and single acts have been condensed and transformed into a label for a conceptualization that cannot be decomposed. Moreover, these three method concepts may be regarded as examples of three scientific approaches which very likely would not have emerged by means of other methods of analysis.

The pattern type analysed has not caused any difficulties in the coding, neither in its single form nor in its compound variants.

The second pattern group consists of a pattern type including problem and two demarcations (30 + 33 + 34), which means a pattern representing titles with implicit intentionality and second-degree extensionality.

Enkätsvar	från klasslärare	(34)
(Questionnaire responses	from school teachers	
30	33	
och klinikföreståndare i	årskurs 3	
and clinic directors	in grade 3)	
33	34	

```
 Pedagogiska problem  vid undervisning  av          (35)
(Pedagogical problems in  the education of
 30                   33                34

 särskoleelever
 disabled pupils)

 Vägen   till och genom   gymnasiet               (36)
(The way to  and through  the gymnasium (grades 10-12)
 30      33

 i  Sverige
 in Sweden)
 34
```

The first example (34) reflects a common activity among the researchers, i.e. the use of questionnaires. In this case, however, the questionnaires themselves are the objective of the report, i.e. the responses, which have been carefully specified. Example (35) focuses on educational problems, not on the pupils themselves. *Särskoleelever* (disabled pupils) is used in this title to specify the problem. Example (36) emphasizes the "pathway" followed in Swedish "gymnasial" (upper secondary school) studies.

This pattern type displays greater complexity than the one discussed earlier, implying that more rules are activated in the automatic coding (see Box 4, Chapter 4.3) and also that there is a risk of misinterpretation. For example, rule 13 has operated in title (35). When the problem has been coded after *vid* (in) has operated, *av* (of) cannot point to a problem. The concept demarcated by *av* (of) instead becomes subordinate to the first extension. In title (36) a stop rule has operated. The coordination of two determinations through a combination of two prepositions is in the coding procedure performed in such a way that the first becomes a "pointer" (according to rule 5).

In this pattern group the procedure has resulted in only one coding error. Consider the following example:

```
 Det fria tillvalet      på grundskolans hög-        (37)
(The free subject choice in upper comprehensive
 30                      33

 stadium och vägen   till  gymnasiet
 school  and the way to    the gymnasium)
         33(30)      34(33)
```

The correct code numbers are given within parentheses. This title expresses two conceptualizations, connected by a conjunction. They are of the same type, i.e. 30 + 33. The coding process goes from left to right and the algorithm has not been able to handle this. The second conceptualization is not coded before *och* (and) operates. *Till* (to) is then ordered according to the preceding concept and coordinated with the part preceding *och* (and). This title may be compared with a correctly coded title which also shows relatively great structural complexity:

En	studie	av	kreativitetsutvecklingen	inom	(38)
(A	study	of	creativity development	within	
40		30		33	

årskurserna 4-9	samt	en undersökning	av
grades 4-9	and [also]	an investigation	of
	40		30

kreativitetens	samvariation	med	intelligens
creativity's	intercorrelation	with	intelligence)

Rule 10, which states that only one main preposition of the *av* (of) type can be present in a clause, here defines "clause" adequately, preventing *samt* (and [also]) from being coded as a connector between *årskurserna* (grades) and *undersökning* (investigation). The conceptualization after *samt* (and [also]) contains a main preposition, which according to rule 12 operates backwards (40 before 30).

The third pattern group is exemplified through the most common structure of explicitly stated intentionality, expanded with first-degree extensionality, that is the 40 + 30 + 33 pattern. The type can be realized as follows:

Mätningar	av	språkfärdighet	i	tyska	(39)
(Measurements	of	language proficiency	in	German)	
40	30		33		

Två	notiser	om	mätning	av	förändring	(40)
(Two	notes	on	the measurement	of	change)	
40		30		33		

Urval	av	elever	till	teoretisk	utbildning	(41)
(Selection	of	pupils	to	theoretical	education)	
40	30		33			

An interesting comparison can be made between (39) and (40). In (40) *mätning* (measurement) is the objective of the study and does not indicate the activity itself, as opposed to (39). Title (40) may be interpreted as if the problem *mätning av förändring* (measurement of change) has a precise meaning for the author. Thus *förändring* (change) is not the problem; instead, *av* (of) has been "degraded" by *om* (on), i.e. a problem component has already been determined before *av* (of) operates. Contrary to the author of "Reflexioner om ..." (Reflections on ...) discussed above, the author of (40) gives the impression of having a well-defined problem area to deal with. Notes are common as a form of presentation when the content does not refer to an empirical investigation. This is, however, the case in (39), which suggests that the author reports on *mätningar* (measurements) that have been performed. Title (40), by contrast, indicates that the author's aim is to discuss certain aspects of measurement; he need not have performed any measurements himself.

Title (41) gives an example of further activities among researchers and/or educational policy makers, namely the development and testing of selection techniques. The example is also interesting in that *teoretisk utbildning* (theoretical education) determines or governs *elever* (pupils) and not *urval* (selection), which it might seem to do at first sight. The title should instead be interpreted as "urval av sådana elever som är lämpliga att *till*höra den grupp som går i teoretisk utbildning" (selection of such pupils as are qualified to belong *to* a group participating in theoretical education). In this case the preposition *till* (to) has been found to be the most adequate to express the function of "assignment to". If the title had instead been worded "Urval av elever *för* teoretisk utbildning" (Selection of pupils *for* theoretical education), education would have been the goal of the selection, i.e. the pupils would be expected to educate themselves in theoretical subjects. No such expectation is conveyed by *till* (to).

There are no unsatisfactory coding results to be reported within this pattern type.

As shown by Table 3 in the previous chapter, the pattern discussed is typical of research reports. However, reports are so frequent in the material that other pattern types also appear in them. Therefore, some examples will also be given of titles showing still greater structural complexity.

Användning	av	ITV	vid	undervisning	i	muntlig	(42)
(Use	of	ITV	at	instruction	in	oral	
40		30	33		34		
framställning							
presentation)							

Studier	av	sociala	relationer	mellan	barn	i	(43)
(Studies	of	social	relations	between	children	in	
40		30		33		34	
folkskoleklasser							
elementary school classes)							

No coding problems have emerged within these pattern types. The "concentric principle" can be found to be at work in both (42) and (43), i.e. the outermost extension demarcates the nearest inward concept.

An example of a title with a high degree of complexity which has been correctly coded is:

Två	utredningar	om	relationerna		(44)
(Two	investigations	concerning	the relations		
40		30			
mellan	brukare,	förvaltare	och	byggare	
between	users,	administrators	and	constructors	
33		33	33		
med	särskilt	avseende	på barn	och ungdom	
with	special	reference	to children	and young people)	
34				34	

This is a research report (not a Swedish "utredning" in the form of an Official Government Report). Investigating is, however, a variant of research activity, and thus *utredningar* (investigations) is coded as being a method. Irrespective of whether the report is interpreted as being a kind of official investigation or a presentation of other investigations, it is representative of the differentiated activities within this group of researchers. Further, the example shows that the connective functions have not increased coding difficulties and also that a multi-word preposition has operated correctly.

Within this type a couple of questionable codings have appeared, namely

Promemoria	rörande	ett	forskningsprojekt	(45)
(Memorandum	concerning	a	research project)	
40	30			

angående	generationsmotsättningar	och	upptagande	av
about	generation gaps	and	adoption	of
33		40(33)		30(34)

vuxenrollen
the adult role)

Preliminär	redovisning	av	resultat	från	en	(46)
(Preliminary	account	of	results	from	a	
40		30		33		

nordisk	utprövning	av	studiematerial	och	näringslivs-
Nordic	test	of	study material	and	economic
		34		34(30)	

synpunkter	på	innehållet	i	dessa
views	on the	content	in	these)
	35(33)		36(34)	

The first example may be compared with (38) above. The coding has been processed in the same way after *och* (and), but here it is not a disconnector. The correct coding within parantheses shows that two conceptualizations are not present. The algorithm does not function when the patterns on both sides of a connector are not in balance (in this case the elements are of different kinds).

Example (46) is, on the whole, characterized by rather a high degree of imbalance. It connects, by right coding (see parentheses), two conceptualizations of different kinds, namely 40 + 30 + 33 + 34 and 30 + 33 + 34. The problem concerning the second conceptualization seems to be the result of an ambiguous abstraction. Moreover, the use of pronouns leads to special difficulties in conceptual analysis, regardless of correct coding.

Finally, the last two incorrect codings will be presented. They belong within a more complex pattern type than the ones just discussed, since they represent a higher degree of intentionality.

Redogörelse	för	mätningar	av	samband	(47)
(Account	of	measurements	of	relatedness	
40			30		

mellan	uppförande	respektive	ordningsbetyg
between	behaviour marks	respectively	discipline marks
33		33	

och ämnesbetyg	för elever	på högstadiet
and subject marks	for pupils	at upper comprehensive
40(33)	70	73

school level)

En jämförelse	mellan	två	system	för	bedömning	(48)
(A comparison	between	two	systems	for	evaluation	
30	33			70		

och betygsättning	av fysikskrivningar	i gymnasiet
and grading	of physics exams	in the gymnasium)
40(70)	30(73)	33

In title (47) the main rule has operated at *för* (for), preventing the balancing rule at *och* (and) from operating. The second example contains one more error, in that *fysikskrivningar* (physics exams) has been assigned a main code, a consequence of *och* (and) not being able to operate as a clause demarcator (according to rule 10).

Within the Goal complex *fysikskrivningar* (physics exams) constitutes the problem, while the report focuses on the comparison between the systems. This double function in the components of the Goal complex may be compared with the title below.

En observationsteknik	för bedömning	av	(49)
(An observation technique	for evaluation	of	
40	70	30	

samarbetskarakteristika	vid	grupparbete
cooperation characteristics	in	group work)
	33	

Only one problem is explicitly stated in this title, but the Problem component in this type of title has a double function, which mirrors the research process very well.

There are a small number of titles of this type. (49) should be interpreted in such a way that an *observationsteknik* (observation technique) is developed in relation to a certain problem, here *samarbetskarakteristika* (cooperation characteristics). In such cases the problem is used instrumentally. Not until the technique has been developed may the *bedömning* (evaluation) of the problem take place.

Further examples of this construction are "Ett system för klassificering av feltyper i diagnostiska skrivprov" (A system for classification of error types in diagnostic written tests), "Ett attitydformulär för studium av elevernas inställning till skolmiljön" (An attitude form for the study of the pupils' attitude to the school environment), and "Forskningsprogram för processanalyser av årskurslöst högstadium" (Research program for process analyses of non-graded upper comprehensive school). The activities denoted as goals by these titles must, seen along the time dimension, be placed after the development of the method. In titles like (48) it would be quite feasible to denote, in a second step, this double function of the *för* (for) structure. But whether and in what manner this is done will be affected by the functional utilization of the register, which to a great extent will be an empirical question.

The five titles presented above as being incorrectly coded make up all the errors in the material tested. Based on the numbers of patterns (n = 871), the proportion is .0057, i.e. no more than six titles out of thousand patterns have been coded incorrectly, due to the inability of the demarcation rules to handle the difference between connection and disconnection.

6.2 Intermediate language functions

With the representational function of language as a starting point, it was assumed in Chapter 3 that there exist different levels of representation as a consequence of the different transformational stages through which documents pass in content description processes. This assumption also implies that different levels of representation are characterized by different degrees of abstraction. In this respect the intermediate language has been defined as the structural representation that should be used in a thesaurus for communication between author and information searcher. The starting point for the generation of this language is the organization of the titles.

The algorithm developed in the present study for the recognition of concepts is based on assumptions of relations with a higher degree of formalization than natural language, i.e. the algorithm is based on intermediate language functions. The concepts are assumed to be part of cognitive structures which, because of the degree of formalization in the titles, are possible to code automatically. The functional relations are here signalled by prepositions. In the preceding section some coding results were demonstrated, in which the algorithm has generated unsatisfactory coding proposals in relation to expectations. The logic presupposed by the algorithm did not coincide with the structure of the title in those cases. It is the conjunctions that have caused trouble in that the proper discrimination between connection and disconnection has not been made. This is a well-known problem in automatic analyses (see e.g., Woods, 1973). The conjunctions as a word class are logical markers in the natural language, whose task is to connect expressions "of the same type", which could imply parts of a clause or whole clauses (cf. Rules 10 and 14, Chapter 4.3). One problem is to discriminate between these two cases when the marker is the same. In addition, when, as in the present algorithm, the coding of single concepts is demarcated by the organization of the prepositions, there is a counteraction between the logical and the functional coding. This implies that the functional relations, i.e. relations between the concepts have not been distinguished either.

In this section an attempt will be made to examine whether the representation of the concepts in a title corresponds to the expected degree of abstraction. This discussion concentrates on structural relations, i.e. relations within concepts.

The concepts determined by prepositions and conjunctions are assumed to have such a construction at the intermediate level that they may, without their context, become functional entities in a register and represent the title from which they are generated. From this point of view certain characteristic features in natural language cannot be accepted, such as inference and reference. Relations in natural and intermediate structures are expressed in different ways. A natural language is more concrete than the intermediate variety. This basic difference may be compared with the active - passive dimension. An active way of writing is "close to things", i.e. close to what is to be described "here and now", while a passive way of writing increases the distance between the writer and what is to be described. One example of this phenomenon, which was discussed in Chapter 4, is a natural language sentence like "I have analyzed titles" transformed to "An analysis of titles", in which the personal and temporal aspects of the assertion disappear resulting in an abstracted concept of the process involved. There are other more or less abstracted conceptions that might be discerned. Within the context of the functions of the coded units and the operational procedure of the algorithm, some titles will here be discussed.

Tjugosex	års	uppföljning	av	en	grupp	elever,	(50)
(Twentysix	years	follow-up	of	a	group	of pupils,	
40			30				

som avgått	enligt	folkskolestadgans
who dropped out	according to	elementary school regulations
30	33	

paragraf	48
paragraph	48)

This title gives a natural impression. With respect to the kind of pupils that the problem refers to, the second part of the title seems to be a little too long, since there now exists a specific term, *studieavbrytare* (drop-out), to designate this kind of pupil. Here, then, is an example of a concrete level. The generated unit *som avgått* (who dropped out) belongs to the above-mentioned active way of writing. A relative pronominal reference is "close to things" and finite verb forms make for concreteness. As soon as a phenomenon has been studied and defined, and thus incorporated in a certain conceptual structure, it can be communicated in abstracted form. As further examples of this two titles from another author may be examined.

Försöksverksamhet	med	nya	former	för	samarbete	(51)
(Experimental work	with	new	forms	for	collaboration	
40	80			70		

mellan	studerande,	lärare	och	övrig personal	vid
between	students,	teachers	and	other staff	at
73		73	73		74

lärarutbildningsanstalter
teachers' training colleges)

Försöksverksamhet	med	nya samarbetsformer	vid	(52)
(Experimental work	with	new collaboration forms	at	
40	80		83	

lärarhögskolan	i	Malmö
the School of Education	in	Malmö)
	84	

The first example is taken from a mimeographed paper produced in 1969. Its goal is collaboration between different categories of staff, and in order to achieve this some experimentation with *nya former* (new forms), vaguely defined, is carried out. When a research report appears in 1972 (example 52) the "forms" are more sharply defined and the author is able to form the concept *samarbetsformer* (collaboration forms).

Such a relatively simple contraction, from the point of view of language structure, is more difficult to accomplish in example (50). The phrase *paragraf 48-elever* (paragraph 48 pupils) has been used at an intermediate stage.

The reason why the pronoun *som* (who) has not been interpreted as a prepositional *som* (as), in the operational sense, is the comma, serving as a logical marker. The rules do not account for relative pronouns, but this case has nevertheless been correctly coded, despite two structural levels. When this unit, procedurally defined, is assigned to a Problem register, its reference (at least in Swedish) is not clear.

In contrast to the relative pronoun the next title shows two codings of a prepositional *som* (as), which thus function according to expectancies.

Kamratbedömning	som	validitetskriterium	och	(53)
(Peer evaluation	as	validity criterion	and	
30	33		33	

som	medel	att	studera	gruppdymaniken
as	means	to	study	the group dynamics)

The two 33 units are interrelated and assigned to the same register, but the second part is in imbalance compared with the first part. *Validitetskriterium* (validity criterion) is assumed by the author to be a communicative concept, as opposed to *gruppdynamiksobservationsinstrument* (group dynamics observation instrument) into which the second concept could be formalized. The cause of the imbalance is probably that the construction with *att* (to) is used when a more general concept cannot be formed. Such a formation requires unambiguous relations to have been established in the research process, providing the basis for the meaning conveyed. The more explicit level of structuring characterizing this unit may require that the elements should be treated as kinds of elementary units, which calls for other algorithmic analyses than are necessary at an intermediate level (cf. Chapter 4).

Another example

Att	mäta	attityder	till	jämställdhet	(54)
(To	measure	attitudes	to	equality)	
30			33		

is a case where an *att* (to) infinitive has been chosen instead of the verbal noun *mätning* (measurement). The infinitive form should instead be interpreted in such a way that it is part of the problem, i.e. the problem is *att mäta* (measuring, to measure) attitudes and a discussion of how this problem area could be tackled. In this connection it may be noted that the same author uses a construction with *att* (to) having another function, namely in the title

Att	mäta	världsmedborgaransvar		med	(55)
(To	measure	world citizenship	responsibility	with	
40				80	
projektiva test					
projective tests)					

Here it is explicitly stated that a process takes place - the instrument employed is mentioned. The researcher states the approach taken, which does not have to be the case in the title (54). *Att mäta* (measuring, to measure) is in this case (55) more related to a research situation. The title may be interpreted as describing a stage in a construction process. From this point of view the coding algorithm may be said to capture the underlying meaning of the construction.

The last two examples have such a high degree of formalization that a decomposition of the infinitive and the following unit would be possible in a second step, i.e. in such a way that *att mäta* (measuring, to measure) is analysed as method in both cases and the second unit as problem. But since it can be assumed that authors formulate their titles very consciously, especially when the title is highly formalized in other respects, it should be possible for an initial *att* (to) construction to enter a register without causing any difficulty. What the concept in example (55) says is that the methodological activity refers to test construction, which implies that it lies within the same functional domain as, for example, the concept *intelligenstestning* (intelligence testing). *Att mäta attityder* (measuring/to measure attitudes) as a problem concept is more like different kinds of attitudes: the same author has at a later stage used *jämställdhets-attityder* (equality attitudes).

The principal difference between *att mäta* (measuring, to measure) and *mätning* (measurement) in this material seems to be that the verbal noun is used when the method is determined and the problem area is well-defined (cf. example (31) in Chapter 6.1). The infinitive belongs to titles which deal with preliminary stages.

The titles discussed so far are examples of structuring at different levels, which are more or less suited for direct representation of scientific concepts. But as opposed to the 5 titles incorrectly coded due to logical misconception, these are characterized by fewer functional cues, indicating a partly lower transformational level. Thus the algorithm can detect stages in the conceptualization, which have not yet reached the intermediate level of processing. There are also a small number of titles in the material which, according to the above discussion, are not characterized by an intermediate structure.

Some examples:

Vad är pedagogik? (56)
(What is pedagogics?)
30

Lönar sig utbildning? (57)
(Pays education?)
30

The typical feature of these titles is that they are complete sentences. The whole sentence is in these cases assigned to the Problem register, but no difficulties arise in their interpretation. The difference between this type and a highly structured title (which is more abstract) is that the latter type is characterized by a greater distance in relation to the phenomenon dealt with, in comparison with examples (56, 57). Such a correspondence between structure and aspect could be utilized for the development of a structurally adapted algorithmic analysis. As an illustration of the effect resulting when the highly formalized algorithm is applied to a conceptually low-structured text two titles are given from the material, both in a natural language variant, reminescent of exchanges of words in a discourse situation.

Så	länge	vi	har	snus,	knäckebröd	och	(58)
(As	long as	we've	got	snuff,	crisp bread	and	
30					30	30	

fruntimmer,	så	nog	blir	Sverige	försvarat
women,	sure	Sweden		will be	defended)
	30				

Man	kan	ju	faktiskt	få	reda	på	ett		och	(59)
(You	can		actually	get	hold	of	one	thing	or	
30						33			40	

annat	om	tentan,	förstås
another	about	the exam,	you see)
	30		30

The first title would generate terms which would be unacceptable in a thesaurus for this research field. The second generates nonsense elements, whose functional relations cannot be processed by the proposed algorithm. However, the two titles have subtitles of a "normal" kind, which guarantee that the information conveyed can still be stored in the system.

The type of titles discussed in the last section altogether makes up approximately 5 % of the material examined. This means that the coding algorithm and the titles have, on the whole, a common logic concerning the structure of the elements to be organized in the registers. These automatically generated registers form the basis for the determination of the vocabulary in a thesaurus. Thus, well-functioning structures and concepts should be available in the titles.

Titles of scientific documents are characterized by different patterns concerning the presence and sequence of order of the components, in this analysis referred to as structural complexity. This seems to co-vary with type of document (also called representation form). The existence of different structural levels pointed out in this chapter is an indication of conceptual variation in the sense that an explicit "natural" structure reflects a "lower" level of transformation or abstraction.

With this discussion and the analyses presented here serving as a general framework, a final description of what is intermediate in documentary language will be proposed.

In I&D processes the thesaurus serves as a means of communication for both indexing and retrieval purposes. For a means of communication to be called a language it has to have a vocabulary (lexicon) and a system of rules. If only one component exists or if instructions are missing as to how the vocabulary and the rules are to function together, this structure is too low to be called a language. Between such a "non-language" and a highly structured means there are more or less well-functioning language variants. As mentioned in Chapter 3, the vocabulary together with explicitly formulated rules is a basic feature of both natural (NL) and artificial language (AL). The more artificial the language variant is, the more precise definitions and general functions it requires.

The language of the titles that in the present material is to form the basis for the generation of an intermediate language (IL) must lie within AL's functional domain. However, analyses have shown that certain features of NL appear in a few titles, which means that the degree of artificiality in the language of titles may vary. Features of NL may possibly be neutralized with the aid of indexing languages and indexers (cf. examples (58, 59) above, however). If, on the other hand, the title is going to be used in building up a retrieval language, more stringent requirements have to be imposed on terminology and logic.

The cognitive model that the coding mechanism presented in the present work is based on adds a new dimension to the information system, the relations between terms being automatically indexed, i.e., they are recognized in the construction of a retrieval language. Thus indexing and retrieval are dependent on each other. The thesaurus then has a mediating function between these two processes. The title, too, has this intermediate function, and so far as the title's conceptual and structural logic can be used for indexing and retrieval, its language is intermediate.

In order to specify in more concrete terms the intermediate language function of the planned thesaurus for the subject field employed, the following section will present a display from the generated registers with examples of the structure of the vocabulary. The function of the registers will also be demonstrated.

6.3 Generated registers

Based on the coding system, registers with different functions have been built up. The functions refer to the three fundamental components of the model. This means that the units assigned to a certain register have something in common. The common features of the units are defined by their having the same function. That the units have the same function does not imply, however, that they must necessarily be homogeneous in other respects. Obvious examples of this have been demonstrated with the method function, where different aspects of "method" may be distinguished in the form of research methods, teaching methods,

methods for reporting, etc.

The work with the construction of the thesaurus, the next step, will not be further discussed here. But since the content, the function, and the proper inferences of the registers constitute the immediate basis for the development and testing of both the terminology and the retrieval mechanism, the purpose of this last section is to give an impression of what, in principle, the registers contain and how they may function. An authentic display of the computer output and statistical analyses of the generated registers can be found in Bierschenk, Bierschenk & Sternerup-Hansson (1979). A vocabulary study of all registers has been carried out. Out of about 2,100 units some 300 aspects (facets) have been distinguished (I. Bierschenk, 1979). Some information from these examinations will be given.

The Problems (register 30) typical of the field of educational research are of a general kind (discipline-oriented), e.g. *pedagogik* (pedagogics), *psykologi* (psychology), *edukation* (systematic instruction), *fostran* (upbringing). This register also describes subfields with a wide range of meaning, such as *begåvning* (ability), *inlärning* (learning), *intelligens* (intelligence), and *prestation* (performance). Other, more teaching-oriented fields are *språkfärdighet* (language proficiency) and *basfärdigheter* (basic skills). A third problem type concerns problems within research itself, such as *datainsamling* (data collection) and *testning* (testing), etc. As pointed out earlier, the problem orientation is the most typical feature of this material. Often the problems are explicitly demarcated (register 33). Typical extensions are localization in space (countries and places) and in time (ages and grades). Further, the problems are often related to school forms, e.g. *grundskolan* (comprehensive school), *gymnasiet* (gymnasium school), and to subjects of study. More general types, such as *skolorganisation* (school organization) and *samhälle* (society), appear as well. The registers 34-37 contain similar units. For example, they denote school form, levels, subjects of study, and population of investigation, such as *elever* (pupils), *flygförare* (aircraft pilots), and *skolledare* (school principals). These registers have more in common with each other than they have with register 33. The closer to the main register its function lies, the more general terms it contains.

This register complex can now be compared with Instrument and Goal. The instrumental vocabulary concerns mainly different material-method systems, e.g. *programmerad undervisning* (programmed instruction), *tvåspråkiga ordlistor* (bilingual word lists), or instruments employed in data collection, such as *observationer* (observations), *personlighetsschemata* (personality schedules), and *videobandspelare* (videotape recorder). Extensions, when appearing, are here, too, of the localization type. The main aspects in the Goal register refer to persons and educational devices, e.g. *dialekttalande elever* (dialect-speaking pupils), *lärare* (teachers), *föräldrar* (parents), *synsvaga* (visually handicapped [people]), *psykiskt utvecklingshämmade* (mentally retarded [people]), and *lärarutbildning* (teacher training), *skolan* (school), *högre studier* (higher education), respectively. The extensions that exist (register 73) refer mainly to groups of persons, but also to school levels. Localizations are grouped within the other subordinate registers.

The circumstances dealt with in this short summary justify still greater expectations as regards the contents of the registers, with the examples of titles from the material already presented. Finally, a study within register 40 yields the following features.

The activities are to a great extent purely research-oriented, expressed by terms such as *forskning* (research), *analys* (analysis), *design* (design), *urval* (selection), *kartläggning* (mapping), *undersökning* (investigation), and the like. Researchers measure effects, construct tests and questionnaires, make data analyses, etc. But other expressions of activity are also numerous. For example, various things are dealt with and discussed in the form of *funderingar* (reflections) and *några metodiska synpunkter på* (some methodological views on); or suggestions are presented as a *förslag* (proposal) or a *promemoria* (memorandum). Investigations of educational matters and production of textbooks are also common activities, e.g. *utredning* (investigation), *rekrytering* (recruitment), *en handbok* (a handbook), *lästeknik* (study techniques).

Before the functioning of the registers is demonstrated, Box 8 will provide a display from the registers 30, 40, 80, and 70, i.e. the ones that correspond to the main components of the model. The units are listed in the authentically generated form, but without operators.

The structures within the units in the registers will be closely studied in connection with the construction of the thesaurus. It will then be important to discuss the kind of "similarity" that exists between the terms for determination of facets. That discussion, however, will not be considered here.

By now, the content in the register should not need any further comment. Examples with an asterisk demonstrate how identical units have more than one function. *Kartläggning* (mapping) may be a method described in one title, but a goal in another; likewise, *en experimentell studie* (an experimental study) may be both method and problem. What these examples show, as pointed out in the analysis in Chapter 6.1, is that lexical forms need not be associated with the components in the model that correspond to those in a more linguistically oriented analytical paradigm. Words which give a concrete impression of a "thing" may thus function as "verbs". This analysis distinguishes and brings to the fore dimensions a manually performed analysis would not have succeeded in doing, owing to habitually learned conceptions and classifications. This is examplified in Box 9 by the concept *skola* (school), which, as part of various compounds, has a high frequency in the material examined.

Box 8. Register display

*Example of functional dissimilarity

Problems (from register 30)

aggressivitet	(aggressiveness)
allmänbegåvning	(general ability)
begåvning	(ability)
begåvningsreserven	(the ability reserve)
begåvningsurvalet	(the ability selection)*
en experimentell studie*	(an experimental study)
individualism	(individualism)
inlärning	(learning)
instudering	(studying)
intelligens	(intelligence)
intelligensbegreppet	(the intelligence concept)
intelligenskrav	(intelligence requirements)
intelligenskvot	(intelligence quotient)
intelligensstandard	(intelligence standard)
intelligensålder	(mental age)
inåtvändhet	(introspectiveness)
kreativ utveckling	(creative development)
meditation	(meditation)
medvetandet	(the consciousness)
mognande*	(maturation) *
mätning	(measurement)
neuros	(neurosis)
personlighetspsykologiska faktorer	(personality factors)
personlighetsutveckling	(personality development)
självbedömning	(self-evaluation)
självförverkligande	(self-realization)
självständighet	(independence)
specialbegåvning	(special ability)
temperamentslära	(theory of temperament)
ett ungdomspsykologiskt problem	(a juvenile problem)
undergivenhet	(submissiveness)
uthållighet	(tenacity)
den utvecklingshämmades identitetsutveckling	(the identity development of mentally retarded [people])
utåtvändhet	(extrovertness)
vardagsinlärning	(everyday learning)

Box 8. (cont.) *Example of functional dissimilarity

Methods (from register 40)

analys	(analysis)*
bearbetning	(processing)
beskrivning	(description)
design	(design)
diskussion	(discussion)
effektmätning	(measurement of effects)
en empirisk studie	(an empirical study)
erfarenheter	(experiences)
en experimentell studie*	(an experimental study)*
faktoranalys	(factor analysis)
funderingar	(reflections)
försök	(experiment)
försöksverksamhet	(experimental work)
granskning	(examination)
en hypotesprövande undersökning	(a hypothesis-testing investigation)
intelligenstestning	(intelligence testing)
kartläggning*	(mapping)*
klassificering*	(classification)*
konstruktion	(construction)
kvantitativa studier	(quantitative studies)
en longitudinell studie	(a longitudinal study)
metodutprövning	(testing of methods)
mätning*	(measurement)*
psykologiska undersökningar	(psychological examinations)
reflexioner	(reflections)
resultatredovisning	(account of results)
sammanställning	(compilation)
skola*	(school)*
standardisering	(standardization)
studier	(studies)
testkonstruktion	(test construction)
en uppföljning	(a follow-up)
upplevelse	(experience)
urval	(selection)
utvärdering	(evaluation)

Box 8. (cont.)

*Example of functional dissimilarity

Instruments (from register 80)	
alkoholskadade föräldrar	(alcoholic parents)
critical incident-metoden	(the critical incidence method)
engelskundervisning	(English language teaching)
grundläggande matematik	(basic mathematics)
institutionsdemokrati	(departmental democracy)
intervjuer	(interviews)
läromedelsframställning	(construction of teaching materials)
några modeller	(some models)
observationer	(observations)
projektiva test	(projective tests)
psykoterapi	(psychotherapy)
skolan*	(the school)*
skolklinik	(school clinic)
två typer av inlärningsmaterial	(two types of learning materials)
utbildning	(education)

Goals (from register 70)	
analys*	(analysis)*
anpassade	(well-adapted [people])
bedömning	(assessment)
dialekttalande elever	(dialect-speaking pupils)
elektronisk databehandling	(electronic data processing)
elever	(pupils)
fackskolan	(professional training school)
föräldrar	(parents)
gymnasiala skolor	(gymnasial schools)
högstadiet	(the upper comprehensive school)
kartläggning*	(mapping)*
klassificering*	(classification)*
lärare	(teachers)
lärarkandidater	(teacher trainees)
psykologer	(psychologists)
psykologutbildningen	(psychologist education)
registrering	(registration)
samarbete	(collaboration)
skolan*	(the school)*
studenter	(students)
synsvaga	(visually handicapped) [people])
tvåspråkiga elever	(bilingual pupils)
ungdom	(young people)
utvecklingsstörda	(mentally retarded [people])
vuxna	(adults)

Box 9. Exemplification of concepts in functional registers

Skolorganisation (School organization) (30)

skolan	(the school)
skolans socialisation	(the socialization of the school)
skolans utveckling	(the development of the school)
skolans kris	(the crisis of the school)
skolans sociologi	(the sociology of the school)
skolnivå	(school level)
skolsegregation	(school segregation)

Skolform/skolstadium (School form/school level) (70)

skolan	(the school)
gymnasiala skolor	(gymnasial schools)
skolväsendets utveckling	(development of the school system)
grundskolans mellanstadium	(intermediate school)
de första årens räkneundervisning	(mathematics instruction in the first school years)
Fackskolan 2	(Professional Training School 2)

Skolform (School form) (33)

skolan	(the school)
grundskolan	(the comprehensive school)
fackskola	(professional training school)
gymnasieskola	(gymnasium)
den svenska enhetsskolan	(the Swedish comprehensive school)
förskola	(nursery school)
den obligatoriska skolan	(the compulsory school)

Undervisningsorganisation (Instructional organization) (80)

skolan	(the school)
skolklinik	(school clinic)

Utbildningsmedel (Educational means) (40)

skola	(school)

Skola (school) as a problem is a subject of research, development, debate, and change. All this is expressed in the Problem register *skolorganisation* (school organization). When *skola* (school) functions as an extension of a problem, it is the school form that is expressed. The school form is often the goal of an activity as well. The last two examples from the Goal register illustrate a textbook aspect of *skola* (school). In an instrumental function, *skola* (school) is an aid; this function may be regarded as the social aspect of the school.

Finally, it should be stressed that in fact the concept *skola* (school) also functions as method. In this sense, the school may be seen as a strategy by means of which one whishes to bring about a change in society. The Method component is here given a broader meaning, since the school may also be seen as an instrument. Method and instrument are components which can form method(instrument)-goal hierarchies in relation to the degree of complexity in the desired goals. In order to reduce method and instrument to a single concept, the term "means" is used.

Consider this title, written long before 1980:

Skola för 80-talet (60)
(School for the 80's)

40 70

In the light of the above discussion and analysis, the means-goal relation expressed in this title probably reflects what the author wants to communicate. Knowledge of the author's activities and field of inquiry in Swedish educational research can validate the interpretation proposed.

7. CONCLUSIONS

Obtaining access to information is regarded as an ever-growing problem by an increasing number of people. Moreover, it seems that information is becoming more and more abstract. In view of this it is of great importance that methods and techniques can be developed which pave the way for a user-oriented organization and re-organization of information.

In the development within information science and related disciplines there is a trend away from systems based on thinking in terms of classes, towards systems building on thinking in terms of functions. This change requires explicitly formulated cognitive models. Otherwise it will hardly be possible to attain the goal of re-organizing information.

In the use of modern I&D systems, the thesauri will be of central importance for communication between the producer and the user of information. The thesauri have an intermediate function, which means that their language structure has to be focused on.

The first description of a document that the user of an I&D system gets is the title, which is often the only description of the document. For this reason, titles have constituted the point of departure in the present study concerning the development of a method for the generation of an intermediate language. The title is supposed to be a condensed version of many empirical observations as described in a document. For the purpose of analysing the relations between the components in a title, prepositions have been used. To be able to use the prepositions as pointers to concepts and conceptual relations, it is of basic importance that their ambiguity at this abstract level can be eliminated.

In the analysis two kinds of prepositions can be distinguished: prepositions referring to intentions, and prepositions referring to extensions. The result of the analysis and the conclusion that prepositions have precise organizing functions have led to the development of an algorithm which makes possible a conceptual coding of titles and the generation of registers with functionally defined content.

An analysis of the structure of titles shows that patterns can be detected that are typical of certain types of documents and less typical of certain other types. The so-called type patterns are characterized by different degrees of structural complexity. This analysis of structure, moreover, can be validated by means of responses taken from an interview study.

The algorithmic analysis of the concepts and the conceptual relations in the titles show that a title which is characterized by structural assymmetry and which contains elements belonging to different levels of abstraction, seems to be in a state of imbalance. This imbalance leads to both structural and conceptual misinterpretation. Finally, it should be pointed out that the algorithm identifies scientific concepts and assigns them to the registers in a functionally more adequate way than would probably have been achieved by a manual analysis.

The present analysis may be regarded as an initial attempt to study a problem area that could be described as "the re-cognition of highly abstracted information". A continuation of this line of research would have important implications for the organization and dissemination of information. It can be expected that in the decade to come, people will no longer ask only for the kind of information service currently available. Rather, what we will probably ask for are "solutions to problems". This development will confront information scientists, especially those who are oriented towards language processing, with new problems. It is most likely that the future will focus on such linguistically based analyses of structure as are aimed at simplifying the representation of abstracted information.

8. SUMMARY

When the purpose of some research activity is to develop methods and techniques for a systematic analysis of language, as regards what information is to be conveyed, questions dealing with the structure assumed to characterize a message receive special emphasis. In view of this it is of considerable importance that a theory can be developed which can explain cognitive phenomena as they appear through language.

In this work attention has been focused on the development of a method and a technique that make possible the mapping of the cognitive structure through which the author of a title of a scientific report communicates what the report is meant to contain. The development of a model and a theory for such a purpose emerges as the basic problem for modern information and documentation (I&D) systems. A dynamic structure should be the goal in the construction of the mechanisms on which I&D systems are based,since information is characterized by the structuring and re-structuring of data, thus being subject to constant change.

There are different forms of representation with different goals. A well-known and common type is made up of hierarchically constructed classification systems, implemented in libraries. Such systems have only a small potential for quick and easy adaptation to a particular structure which may be employed in a search by looking for new information. Once a report is put on a shelf it is bound to its place. For the creation of a flexible organization, a dynamically functioning space is in a sense provided by facet classification. In such systems there are greater possibilities for lateral relations.

Recent efforts to structure information involve networks and schemata. Whereas the former type aims at finding out how many "semantic primitives" are needed for a synthetic

formation of concepts, the latter tries to explore the advantages of an adaptively operating process. Schemata are based on an inferential strategy, which implies that only some or one of the components in the schema model may be activated.

The starting-poing in the present experiment is that an intermediate language can be generated from titles of scientific works, and that such a language can be displayed in a thesaurus. An intermediate language capable of representing the cognitive structure of a message has a higher degree of structuring than those which do not have this capacity. The cognitive structure of a field of science is conveyed through scientific documents. They contain different statements about empirical observations which may also be represented in sentence form, as in the N_1 v N_2 paradigm. If, however, the intentions, i.e. the underlying propositions of sentences, can be denoted, another paradigm is required, a paradigm which accounts for underlying propositions at a somewhat more abstract level. The paradigm adequate to this purpose, provided that the language under consideration belongs to the Indo-European family, is the Agent-action-Object (AaO) paradigm. An aggregation of different observations and the causal relations that scientific reporting is establishing between different phenomena require a paradigm with the capacity to represent the scientific process and which, therefore, has to be even more abstract than the AaO paradigm.

The fundamental components in a research process are "problem", "method" and "goal". A paradigm which can express the result of a number of abstracted propositions is the Problem-method-Goal paradigm (PmG). To study a title from the point of view of this paradigm builds on the assumption that there are cues in the overt structure of the title which indicate that the components represent the author's conceptualization. In order to analyse the relations between the single components in the title, the organizing function of the prepositions has been utilized in this work. The important advantage in the use of prepositions for a conceptual analysis of titles is that, in principle, they function unambiguously on the title level compared to the natural (concrete) language level. The concepts are related with respect to their intentions or extensions, where such labels as localization and direction are generalized.

To make possible automatic coding of concepts and conceptual relations, a set of rules has been formulated. According to this set, a title consists of one or more conceptualizations, and so demarcation rules are necessary. Certain editings in the empirical material have been made, e.g. by inserting demarcation markers in the form of commas to indicate connection, and dashes and full stops to mark disconnection. The basic principle in the rule systems is that the prepositions point forwards, which entails an algorithmic consequence, namely that when all elements after each intentional preposition have been determined, only the method is left. 14 rules operate by means of a dictionary consisting of 39 operators, 12 stop phrases, and 5 conjunctions. The program is written in ASCII-FORTRAN.

The algorithm has been tested on a data base which is representative of Swedish educational research. The data base includes bibliographic descriptions, mainly according to the APA standard. For the evaluation a total of about 9,000 bibliographic descriptions of works written in Swedish have been used.

According to the analytical model employed, the conceptualizations may be extended to varying degrees. The most extended case is represented by a pattern which activates a Problem component, a Method component, an Instrument component, and a Goal component, together with possible extensions. Thus a pattern is a structural representation of a conceptualization. The most characteristic feature in the pattern shows that the Problem component appears single, connectively, or together with one extension. The more complex the patterns are, the less often they appear. This has been reflected in a directed graph (Chapter 5.2). There are links between Method and Problem, as well as within the Problem complex. Moreover, Instrument "governs" Goal to a greater extent than

Method "governs" Instrument and Goal. This picture of the research process within educational research is in line with the problem orientation that the authors of the titles have expressed in an interview study.

On the basis of the representational function of language, it has been proposed here that the different levels of abstraction resulting from varying transformational stages are a consequence of conceptual development. To generate an intermediate language which characterizes a document at a particular level of abstraction requires the structural elements to have the same degree of abstraction (structural level) in one and the same title in order to unambiguously map the conceptual relations that they represent. An automatic conceptual information processing is performed when the formalism of the algorithm and the titles coincide. Thus natural language features invalidate the coding.

Finally, it should be mentioned that the model described and operationalized has been employed in automatic identification and coding of such dimensions that a manual analysis could have performed only with difficulty, since it builds on habituated frames of reference.

9. REFERENCES

Abelson, R.P. The structure of belief systems. In: R.C. Schank & K.M. Colby (Eds.) *Computer models of thought and language*. San Francisco: Freeman, 1973. Pp. 287-339.

Aitchison, J. The Thesaurofacet: A multipurpose retrieval language tool. *Journal of Documentation*, 1970, *26* (3), 187-203.

Allén, S. (Ed.) *Frequency dictionary of present-day Swedish based on newspaper material. 1. Graphic words. Homograph components*. Stockholm: Almqvist & Wiksell International, 1970.

Allén, S. On phraseology in lexicology. *Cahiers de Lexicologie*, 1976, *29* (2), 83-90.

Allén, S. Text-based lexicography and algorithmic text analysis. *ALLC Bulletin*, 1977, *5* (2), 126-131.

Allén, S., Berg, S., Järborg, J., Löfström, J., Ralph, B. & Sjögreen, C. *Frequency dictionary of present-day Swedish based on newspaper material. 4. Morphemes. Meanings*. Stockholm: Almqvist & Wiksell International, 1980.

American Psychological Association (APA). The role of the technical report in the dissemination of scientific information. Project on scientific information exchange in psychology. Washington: *APA-PSIEP report*, No. 13, 1965.

Anderla, G. *Information in 1985. A forecasting study of information needs and resources*. Paris: OECD, 1973.

Artandi, S. Document description and representation. *Annual Review of Information Science and Technology*, 1970, 5, 143-169.

Austin, D. The development of PRECIS, and introduction to its syntax. In: H. Wellisch (Ed.) *The PRECIS index system*. New York: Wilson, 1977. Pp. 3-28.

Bartlett, F.C. *Remembering. A study in experimental and social psychology*. London: Cambridge University Press, 1932.

Bierschenk, B. *Datorbaserad litteratursökning. /Computer-based search for literature./* Lund: Studentlitteratur, 1973. /In Swedish/

Bierschenk, B. En modell för ett interaktivt informations- och dokumentationssystem. /A model for an interactive information and documentation system./ *Pedagogisk dokumentation*. (Malmö: Department of Educational & Psychological Research, No. 26, 1974. (a) /In Swedish/

Bierschenk, B. Perception, strukturering och precisering av pedagogiska och psykologiska forskningsproblem på pedagogiska institutioner i Sverige. /Perception, structuring and definition of educational and psychological research problems on departments of educations research in Sweden./ *Pedagogisk-psykologiska problem.* (Malmö: Department of Educational & Psychological Research), No. 254, 1974. (b) /In Swedish/

Bierschenk, B. A new approach to psychometric problems in the analysis of pre-numeric data. *Didakometry.* (Malmö: Department of Educational & Psychological Research), No. 55, 1977. (a)

Bierschenk, B. Research planning from a micro-ecological perspective: Summary of interview study. *Educational and Psychological Interactions.* (Malmö: Department of Educational & Psychological Research), No. 60, 1977. (b)

Bierschenk, B. Innehållsanalys som forskningsmetod. /Content analysis as a method of research./ *Kompendieserien.* (Malmö: Department of Educational & Psychological Research), No. 25, 1978. (a) /In Swedish/

Bierschenk, B. *Simulating strategies of interactive behavior.* (Studia Psychologica et Paedagogia, 38.) Lund: Gleerup, 1978. (b)

Bierschenk, B. En longitudinell analys av kunskapsutvecklingen inom utbildningsforskningen. /A longitudinal analysis of knowledge development within educations research./ *Pedagogisk-psykologiska problem.* (Malmö: Department of Educational & Psychological Research), No. 355, 1979. /In Swedish/

Bierschenk, B. & Bierschenk, I. *A system for a computer-based content analysis of interview data.* (Studia Psychologica et Paedagogica, 32.) Lund: Gleerup, 1976.

Bierschenk, B., Bierschenk, I. & Sternerup-Hansson, A. Ett datorprogram för syntaktisk kodning av titlar till vetenskapliga skrifter. /A computer program for syntactic coding of titles of scientific documents./ *Testkonstruktion och testdata.* (Malmö: Department of Educational & Psychological Research), No. 36, 1979. /In Swedish/

Bierschenk, I. Datorbaserad innehållsanalys: Kodningsmanual. /Computer-based content analysis: Coding manual./ *Pedagogisk dokumentation.* (Malmö: Department of Educational & Psychological Research), No. 52, 1977. /In Swedish/

Bierschenk, I. Försök med automatisk separering av referenser i en flerspråkig databas. /Experiments with automatic separation of references in a multi-lingual data base./ *Testkonstruktion och testdata.* (Malmö: Department of Educational & Psychological Research), No. 34, 1978. /In Swedish/

Bierschenk, I. Trunkerade register. Preliminär konstruktion. /Truncated registers. Preliminary construction./ Mimeographed. (Malmö: Department of Educational & Psychological Research), June, 1979. /In Swedish/

Bierschenk, I. Intermediate language structure. Doctoral Dissertation (Göteborg: Department of Computational Linguistics), 1980. Microfilms, No. 81-70, 037.

Bobrow, D.G., Fraser, J.B. & Quillian, M.R. Automated language processing. *Annual Review of Information Science and Technology*, 1967, *2*, 161-186.

Braun, S. & Schwind, C. Automatic, semantic-based indexing of natural language texts for information retrieval systems. *Report*. (München: Technische Universität, Institut für Informatik), No. 7505, 1975.

Brodda, B. (K)overta Kasus i svenskan. /(C)overt cases in Swedish./ *PILUS*. (Stockholm: Department of Linguistics), No. 18, 1973. /In Swedish/

Brown, R. *A first language. The early stages*. Cambridge: Harvard University Press, 1973.

Browning, D.C. (Ed.) *Everyman's thesaurus of English words and phrases*. London: Pan Books, 1971.

Bunge, M. *Scientific research I. The search for system*. New York: Springer, 1967.

Cedvall, M. Semantisk analys av processbeskrivningar i naturligt språk. /Semantic analysis of process descriptions in natural language./ *Linköping Studies in Science and Technology. Dissertations*, 18. (Linköping: Department of Mathematics), 1977. /In Swedish/

Cofer, C.N. (Ed.) *The structure of human memory*. San Francisco: Freeman, 1976.

Coyaud, M. *Introduction à l'étude des langages documentaires*. Paris: Klincksieck, 1966.

Coyaud, M. & Siot-Decauville, N. *L'analyse automatique des documents*. Paris: Mouton, 1967.

Damerau, F.J. Automated language processing. *Annual Review of Information Science and Technology*, 1976, *11*, 107-161.

van Dijk, T.A. Perspective paper: Complex semantic processing. In: D. Walker, H. Karlgren & M. Kay (Eds.) *Natural language in information science*. Stockholm: Skriptor, 1977. Pp. 127-163.

Educational Resources Information Center (ERIC). *Thesaurus of ERIC descriptors*. New York: CCM Information Corporation, 1975.

EUDISED. *Multilingual thesaurus*. Paris: Council of Europe, 1973.

Fairthorne, R.A. Content analysis, specification, and control. *Annual Review of Information Science and Technology*, 1969, *4*, 73-109.

Faught, W.S. Motivation and intensionality in a computer simulation model. *Report*. (Stanford University: Stanford AI Laboratory), No. AIM-305, 1977.

Fillmore, C.J. The case for case. In: E. Bach & R.T. Harms (Eds.) *Universals in linguistic theory*. New York: Holt, 1968.

Friedman, J. A computational treatment of case grammar. In: J. Hintikka et al. *Approaches to natural language*. Dortrecht-Holland: Reidel, 1973. Pp. 134-152.

Gilbert, G.N. Measuring the growth of science. A review of indicators of scientific growth. *Scientometrics*, 1978, *1* (1), 9-34.

Greene, J. *Tänkande och språk. /Thought and language: Theoretical and experimental studies./* Stockholm: Wahlström & Widstrand, 1977. /Swed. transl./

Grimm, H. On the child's acquisition of semantic structure underlying the word field of prepositions. *Language and Speech*, 1975, *18* (2), 97-119.

Gross, H. Simple sentences. Comments on Householder's paper. Preprint. To appear in Preceedings from the Nobel Symposium on Text Processing, Stockholm, 1980.

Helbig, G. & Schenkel, W. *Wörterbuch zur Valenz und Distribution deutscher Verben*. Leibzig: VED Verlag Enzyklopädie, 1973.

Hillman, D. & Kasarda, A. The LEADER retrieval system. *Spring Joint Computer Conference*, 1969, *34*, 447-455.

Jernryd, E. Informations- och dokumentationsproblem på en forskningsinstitution: Några synpunkter från explorativa intervjuer med projektforksare. /Information and documentation problems within a research department. Some viewpoints from explorative interviews with grant supported researchers./ *Pedagogisk dokumentation*. (Malmö: Department of Educational & Psychological Research, No. 42, 1976. /In Swedish/

Jernryd, E. Några lärarutbildares syn på FoU-arbetets spridning och återkoppling. /Some teacher educators' view on dissemination and feed-back of RoD work./ *Särtryck och småtryck*. (Malmö: Department of Educational & Psychological Research), No. 241, 1978. /In Swedish/

Karlgren, H. Homeosema prepositionsuttryck. /Homeosemic prepositional phrases./ In: H. Karlgren (Ed.) *Homeosemi*. Stockholm: Skriptor, 1974. Pp. 75-115. /In Swedish/

Karlgren, H. Challenge paper: Homeosemy - On the linguistics of information retrieval. In: D. Walker, H. Karlgren & M. Kay (Eds.) *Natural language in information science*. Stockholm: Skriptor, 1977. Pp. 167-181.

Kintsch, W. *The representation of meaning in memory*. New York: Wiley, 1974.

Klein, S. & Simmons, R.F. A computational approach to grammatical coding of English words. *Journal of the Association for Computing Machinery*, 1963, *10* (July), 334-347.

Krippendorff, K. Models of messages: Three prototypes. In: G. Gerbner et al. (Eds.) *The analysis of communication content*. New York: Wiley, 1969. Pp. 69-106.

Lewis, D. General semantics. In: D. Davidson & G. Harman (Eds.) *Semantics of natural language.* Dordrecht-Holland: Reidel, 1972. Pp. 169-218.

Miller, G.A. & Johnson-Laird, P.N. *Language and perception.* London: Cambridge University Press, 1976.

Mølgaard-Hansen, R. UDC, DC, and LC in competitions on the domain of the university library. *Tidskrift för dokumentation,* 1968, *24* (1), 1-7.

O'Connor, J. Text searching retrieval of answer-sentences and other answer-passages. *Journal of the American Society for Information Science,* 1973, *24* (6), 445-460.

Oller, J.W. & Sales, B.D. Conceptual restrictions on English: A psycholinguistic study. *Lingua,* 1969, *23,* 209-232.

Osgood, C.E., Suci, G.J. & Tannenbaum, P.H. *The measurement of meaning.* Urbana: University of Illinois Press, 1957.

Piaget, J. *The origins of intelligence in children.* New York: Norton, 1963.

Piaget, J. & Inhelder, B. *The child's conception of space.* London: Routledge & Kegan Paul, 1956.

Price, de Sola, D.J. *Little science, big science.* Columbia: University Press, 1963.

Quillian, R. Semantic memory. In: M. Minsky (Ed.) *Semantic information processing.* Cambridge: MIT Press, 1968. Pp. 216-270.

Ralph, B. Prepositioners presuppositioner: Några lokativiska exempel. /Presuppositions of prepositions. Some locative examples./ Mimeographed. (Göteborg: Department of Computational Linguistics), 1977. /In Swedish/

Ranganathan, S.R. *Colon classifications: Paper prepared for the Rutgers Seminars on systems for the intellectual organization of information.* New Brunswick: Rutgers, 1964.

Reid, L.S. Toward a grammar of the image. *Psychological Bulletin,* 1974, *81* (6), 319-334.

Richmond, P. Document description and representation. *Annual Review of Information Science and Technology,* 1972, *7,* 73-102.

Robinson, D. Indexing nonbook materials by PRECIS. In: H. Wellisch (Ed.) *The PRECIS index system,* 1977. Pp. 169-174.

Sager, N. Perspective paper: Computational linguistics. In: D. Walker, H. Karlgren & M. Kay (Eds.) *Natural language in information science.* Stockholm: Skriptor, 1977. Pp. 75-100.

Salton, G. The identification of document content: A problem in automatic information retrieval. In: *Proceedings of a Harvard symposium on digital computers and their applications.* Cambridge: Harvard University Press, 1962.

Salton, G. Automated language processing. *Annual Review of Information Science and Technology*, 1968, *3*, 169-199.

Salton, G. (Ed.) *The SMART retrieval system. Experiments in automatic document processing.* Englewood Cliffs: Prentice-Hall, 1971.

Schank, R.C. Conceptual dependency: A theory of natural language understanding. *Cognitive Psychology*, 1972, *3* (4), 552-631.

Schank, R.C. Identification of conceptualizations underlying natural language. In: R.C. Schank & K.M. Colby (Eds.) *Computer models of thought and language.* San Francisco: Freeman, 1973. Pp. 187-247.

Shapiro, S.C. & Kwasny, S.C. Interactive consulting via natural language. *Communications of the Association for Computing Machinery*, 1975, *18* (8), 459-462.

Sharp, J.R. Content analysis, specification and control. *Annual Review of Information Science and Technology*, 1967, *2*, 87-122.

Simmons, R.F. Semantic networks: Their computation and use for understanding English sentences. In: R.C. Schank & K.M. Colby (Eds.) *Computer models of thought and language.* San Francisco: Freeman, 1973. Pp. 63-113.

Soergel, D. *Indexing languages and thesauri: construction and maintenance.* Los Angeles: Melville, 1974.

Sparck Jones, K. & Kay, M. *Linguistics and information science.* New York: Academic Press, 1973.

Sparck Jones, K. & Kay, M. Linguistics and information science: A postscript. In: D. Walker, H. Karlgren & M. Kay (Eds.) *Natural language in information science.* Stockholm: Skriptor, 1977. Pp. 183-192.

Stone, P., Dunphy, D., Smith, M. & Ogilvie, D. *The general inquirer: A computer approach to content analysis.* Cambridge: MIT Press, 1966.

Taulbee, O. Content analysis, specification, and control. *Annual Review of Information Science and Technology*, 1968, *3*, 105-136.

Universal decimal classification (UDC). (Swedish shortened edition.) Stockholm: Tekniska litteratursällskapet, 1961.

Wearing, A.J. Remembering complex sentences. *Quarterly Journal of Experimental Psychology*, 1972, *24*, 77-86.

Welin, C.W. Strukturell flertydighet hos kedjor av prepositionsuttryck. /Structural ambiguity in chains of prepositional expressions./ In: H. Karlgren (Ed.) *Homeosemi.* Stockholm: Skriptor, 1974. Pp. 115-153. /In Swedish/

Wellisch, H. (Ed.) *The PRECIS index system.* New York: Wilson, 1977.

Werner, H. & Kaplan, B. *Symbol formation: An organismic-developmental approach to language and the expression of thought.* New York: Wiley, 1963.

Wersig, G. & Neveling, U. (Eds.) *Terminology of documentation.* Paris: Unesco, 1976.

Wilks, Y. An artificial intelligence approach to machine translation. In: R.C. Schank & K.M. Colby (Eds.) *Computer models of thought and language.* San Francisco: Freeman, 1973.

Winograd, T. *Understanding natural language.* New York: Academic Press, 1972.

Winston, P.H. (Ed.) *The psychology of computer vision.* New York: McGraw Hill, 1975.

Woods, W.A. An experimental parsing system for transition network grammars. In: R. Rustin (Ed.) *Natural language processing.* New York: Algorithmics Press, 1973. Pp. 111-154.

Department of
Educational and
Psychological Research
School of Education
Lund University
Malmö, Sweden

Abstract card

Refecence card

Bierschenk, I. An information processing experiment. A method for the generation of an intermediate language for representing scientific information. Didakometry (Malmö, Sweden: Department of Educational & Psychological Research), No. 62, 1981.

This experiment starts with the assumption that the structure and representation of scientific information should correspond to the cognitive structure assumed to exist in both user and producer of information. The model of investigation of cognitive representation is based on overt manifestations of concepts and conceptual relations as they emerge in the abstract language of titles of scientific documents. On the basis of this language structure an algorithm has been developed and tested using the cue function of prepositions for automatic conceptual coding. The relevance of the concepts is judged with respect to a schema model containing some basic components of research itself. The algorithm generates the assumed scientific concepts and assigns them to different data registers.

Keywords: Cognitive science, computational linguistics, information processing, information science, linguistics, representation language, text processing, thesaurus.

Bierschenk, I. An information processing experiment. A method for the generation of an intermediate language for representing scientific information. Didakometry (Malmö, Sweden: Department of Educational & Psychological Research), No. 62, 1981.

ISSN 0046-0230